SAMPLING THEORY OF SURVEYS WITH APPLICATIONS

P. V. SUKHATME

SAMPLING THEORY OF SURVEYS WITH APPLICATIONS

Sampling Theory of Surveys with Applications

by

PANDURANG V. SUKHATME
Ph.D., D.Sc.
Chief, Statistics Branch, Economics Division, Food and Agriculture Organization of the United Nations; Formerly Statistical Adviser, Indian Council of Agricultural Research, New Delhi, India.

New Delhi, India . The Indian Society of Agricultural Statistics
Ames, Iowa, U.S.A. . The Iowa State College Press

PRINTED BY THE BANGALORE PRESS, BANGALORE, INDIA

To
Professor Jerzy Neyman

PREFACE

This book is an outgrowth of lectures on sample surveys which the author has delivered since 1945 at the Indian Council of Agricultural Research, subsequently at the International School on Censuses and Statistics in 1949–50 held at Delhi under the auspices of the Food and Agriculture Organization of the United Nations, at the two summer sessions conducted by the Indian Society of Agricultural Statistics in 1950 and 1951, and finally at the Statistical Laboratory of the Iowa State College, Ames, Iowa, U.S.A., in the spring of 1952.

There was no plan at first of publishing a book and the notes prepared for the lectures were mimeographed for the use of the students, but as the scope of the course was gradually enlarged, suggestions were received that the lectures should be published in the form of a text for teaching at colleges and universities. It was felt that this publication would fulfil a real need for a systematic treatment of the sampling theory in relation to large-scale surveys. About the same time the Conference of the Food and Agriculture Organization of the United Nations recommended at its Sixth Session that a book be prepared incorporating a comprehensive treatment of the sampling theory of surveys and its applications so as to be of direct assistance to the sampling experts working in various countries in their efforts to introduce the sampling method for improvement of agricultural statistics. The mimeographed notes were accordingly reorganized and amplified to include illustrative material on agricultural surveys from different countries; the publication of the present book is the result.

In keeping with its objectives the book is primarily designed to serve the needs of a text for teaching an advanced course in sampling theory of surveys and of a reference book for statisticians entrusted with the planning of surveys for collecting statistics. Every attempt has been made to present all the modern developments of sampling theory which are of importance in survey work. Some of the results have already appeared in the papers published in the *Journal of the Indian Society of Agricultural Statistics*. These might appear new to many readers since they might not have seen this Journal. The book also gives a number of results which are being published for the first time. Among these should be mentioned particularly the algebraic treatment of non-sampling errors whose importance relative to sampling errors has not been sufficiently stressed in the literature on the subject.

In order that the theory presented in the book should be of direct assistance in practice, it is illustrated with examples of actual surveys so as to serve the special needs of under-developed countries in the field of sampling, as recommended by FAO. These examples are oriented largely around agricultural statistics, in keeping with the author's experience in this field and FAO's interest, and relate to surveys for the estimation of crop acreage, yield, incidence of insect pests on crops, livestock numbers and their products, other farm facts and fisheries production. The author is conscious that these examples by themselves will not meet the entire needs of sampling workers, particularly those from the economically less developed countries where the resources available for planning surveys are meagre and a large majority of the people are illiterate, do not appreciate the purpose of the inquiry, nor know the correct answers to the questions put to them. The contribution to the total error in the result arising from this latter factor is very large in these countries and emphasizes the great value of developing satisfactory measurement techniques before attempting nation-wide surveys. The relevant theory bearing on this question has been discussed in Chapter X. What is further needed is a simple exposition of a few typical surveys. Such a book is nearing completion and it is hoped to make it available soon.

The need for keeping the volume within reasonable size has prevented any elaborate supporting description of the theory and examples given in the book. The author's aim all along has been to present the theory in as straightforward a manner as possible. The only pre-requisites are college algebra, elements of calculus and principal statistical methods such as those covered in *Statistical Methods for Agricultural Workers*, by V. G. Panse and the author. Even so the author is aware that at places the treatment has become too terse. Such sections have been marked with an asterisk to indicate that the portion can be left over from the first reading without losing the continuity of the text.

The author has received considerable assistance in preparing the book from his former colleagues in India. First of all he gratefully acknowledges the encouragement and help which he received from his former Chief, Mr. P. M. Kharegat, then Secretary to the Ministry of Agriculture, Government of India, to whose farsightedness are principally due the advances which India has made in the field of sampling. He is indebted to Messrs. V. G. Panse, G. R. Seth, K. Kishen, R. D. Narain, O. P. Aggarwal and B. V. Sukhatme who read parts of the manuscript and made numerous suggestions to improve the presentation; to Messrs. K. S. Krishnan, S. H. Ayer and K. V. R. Sastry who worked through the examples; and to Mrs. Evans of the Statistics Branch of FAO who checked through them and also helped in the preparation of the index to the book; to Dr. P. N. Saxena who shouldered a particularly heavy responsibility of reading critically the manuscript and the proofs; and to Suzanne Brunelle and Mary Nakano for their typing and secretarial help. The author also likes to express his thanks to Dr. T. A. Bancroft, Dr. D. J. Thompson and other members of the staff of the Statistical Laboratory, Iowa State College, with whom he had the opportunity to work as visiting professor during the spring term of 1952 and to Marshall Townsend of the Iowa State College Press for their interest and encouragement in the publication of the book. Last but not least the author is indebted to Mr. Norris E. Dodd, Director-General of the FAO, who invited the author to come to FAO to head the Statistics Branch, which gave him the opportunity to appreciate more fully the urgent need for promoting sampling for improving agricultural statistics in under-developed countries; and to Dr. A. H. Boerma, Director of Economics Division of the FAO, for his constant encouragement and advice.

September 1953. PANDURANG V. SUKHATME.

CONTENTS

PAGE

III. STRATIFIED SAMPLING

A. SELECTION WITH EQUAL PROBABILITY

B. VARYING PROBABILITIES OF SELECTION

PAGE

IV. RATIO METHOD OF ESTIMATION

A. Sampling with Equal Probability of Selection

B. Sampling with Varying Probabilities of Selection

V. REGRESSION METHOD OF ESTIMATION

PAGE

VI. CHOICE OF SAMPLING UNIT

A. Equal Clusters

B. Unequal Clusters

PAGE

IX. SYSTEMATIC SAMPLING

X. NON-SAMPLING ERRORS

A. Observational Errors

PAGE

B. INCOMPLETE SAMPLES

LIST OF EXAMPLES

Example Number	Page Number	Technique	Subject
5.1	215	Regression estimate	Survey for estimating livestock numbers, with village area as auxiliary variable
6.1	241	Efficiency of clusters of varying size	Survey for estimating crop acreage (wheat)
*	248	Example of negative intra-class correlation ρ (cluster sampling more efficient than random sampling)	Investigation of proportions of males in a household
6.2	251	Efficiency of cluster sampling compared with simple random sampling	Survey for estimating crop acreage (wheat)
6.3	253	Relationship between variance and cluster size, Smith's law	Size of plot for crop-cutting (wheat)
6.4	256	Relationship between variance and cluster size, Jessen's law	Clusters of fields for estimating area under crop (wheat)
*	258	Efficiencies of various sizes of sampling unit	A general purpose survey for obtaining farm facts
6.5	278	Cluster sampling: efficiency of different procedures of sampling and estimation	Estimation of acreage under crop (wheat)
*	297	Calculation of optimum sub-sampling rate for surveys involving several characters	Survey for acreage under crops (wheat, gram, maize, sugarcane)
7.1	298	Two-stage sample: calculation of optimum sub-sampling rate	Crop-cutting survey for yield (rice)
*	321	Two-stage sampling, unequal units: comparison of alternative estimates	Survey for crop yield (wheat)
*	325	Two-stage sampling, unequal units: efficiency of alternative estimates	Survey for crop yield (wheat, cotton)

* Starred items are illustrative examples not numbered in the text.

INDEX TO PRINCIPAL NOTATION

(The number against each symbol indicates the page where it is first introduced and explained)

SPECIAL SYMBOLS

$C.V.$	Coefficient of variation
Cov	Covariance
C	Cost function
$c_0, c_1, c_2, \ldots$	Coefficients of cost functions
$M.S.E.$	Mean square error
$S.E.$	Standard error
V_1	First approximation to variance
V_2	Second approximation to variance
V_S	Variance of a stratified sample
V_P	Variance of a stratified sample, proportional allocation
V_N	Variance of a stratified sample, Neyman allocation
V_R	Variance of a random sample
V_{US}	Variance of an unstratified sample
V_{Sy}	Variance of a systematic sample

CHAPTER I

BASIC IDEAS IN SAMPLING

1.1 Sampling Method

A sampling method is a method of selecting a fraction of the population in a way that the selected sample represents the population. Everyone of us has had occasion to use it. It is almost instinctive for a person to examine a few articles, preferably from different parts of a lot, before he or she decides to buy it. No particular attention is, however, paid to the method of choosing articles for examination. A wholesale buyer, on the other hand, has to be careful in selecting articles for examination as it is important for him to ensure that the sample of articles selected for examination is typical of the manufactured product lest he should incur in the long run a heavy loss through wrong decision. Similarly, in obtaining information on the average yield of a crop by sampling, it is not sufficient to ensure that the fields to be included in the sample come from different parts of the country, for, the sample may well contain a very much larger (or smaller) proportion of fields of a particular category like irrigated, manured or growing improved variety, than is present in the population. If any category is consistently favoured at the expense of the other, the sample will cease to represent the whole. Even if the sample is selected in such a way that the proportions in the sample under different categories agree with those in the population, the sample may not still represent the population. A sampling method, if it is to provide a sample representative of the population, must be such that all characteristics of the population, including that of variability among units of the population, are reflected in the sample as closely as the size of the sample will permit, so that reliable estimates of the population characters can be formed from the sample.

1.2 Standard Error

Whatever be the method of selection, a sample estimate will inevitably differ from the one that would be obtained from enumerating the complete population with equal care. This difference between the sample estimate and the population value is called the sampling error. The larger the sample, the smaller will obviously be the sampling error on the average and the greater will be our confidence in the results. A sampling method, if it is to be serviceable, must provide some idea of the sampling error in the estimate on an average. We must, for instance, be able to form a precise idea of the extent to which we are likely to be in error on an average in estimating the yield of a crop from the sample. Several measures are available for the purpose. One such measure of the average magnitude of the sampling error is called the standard error of the estimate and provides a measure of the reliability, as it were, of the sample estimate. It is the magnitude of the standard error which will determine whether a sample estimate is useful for a given purpose. This, in turn, will depend upon the break down expected of the results. If, for example, estimates of crop acreages are required for every village of the State and for all crops, major or minor, there will be little point in using the sampling method.

1.3 Principle of Choosing among Alternative Sampling Methods

Practical considerations have also to be kept in view in the use of a sampling method. Crop-cutting surveys for estimating the average yield of a crop provide a good example for illustration. It is not enough in a crop-cutting survey to select a sample of fields representative of the total number under the crop and sample-harvest the selected fields at the time of the visit of the enumerator; it is also necessary to ensure that the selected fields are reached on the dates the cultivators would harvest them. Only then would the distribution of sample-harvesting over time correspond to the distribution over time of actual harvesting. The procedure of sample-harvesting should also correspond, in so far as practicable, to the one adopted by the cultivator so that what is observed would correspond to what is gathered by the cultivator, which is what one wants to estimate. Further, a sampling method, if it is to be acceptable in practice, must be simple, fit into the administrative background and local conditions and ensure the most effective use of the resources available to the sampler. The guiding principle in the choice of a sampling method is, in fact, the principle of securing the desired result with the reliability required at minimum cost or with the maximum reliability at a given cost, making the most effective use of the resources available.

1.4 Probability Sampling

To fulfil the above requirements, it is necessary that the method of sampling be objective, based on laws of chance. The method is called the method of *probability sampling*. In this method, the sample is obtained in successive draws of a unit each with a known probability of selection assigned to each unit of the population at the first draw. At any subsequent draw, the probability of selecting any unit from among the available units at that draw may be either proportional to the probability of selecting it at the first draw or completely independent of it.

The successive draws of a probability sample may be made with or without replacing the units selected in the previous draws. The former is called the procedure of sampling with replacement, the latter without replacement.

The application of the method presumes that the population can be subdivided into distinct and identifiable units called sampling units. The units may be natural units, such as individuals in a human population or fields in a crop-estimating survey or natural aggregates of such units like families or villages; or they may be artificial units, such as a single plant, a row of plants or a plot of size, say, 10×10 square feet in sampling a field of wheat. In general, for a given proportion of the population to be sampled, the smaller the sampling unit the more accurate will be the sample estimate. The application of the method naturally presupposes the availability of a list of all the sampling units in the population. This list is called the *frame* and provides the basis for the actual selection of the sample. An example of a frame is furnished by a list of farms, where one exists, or suitable area-segments like the village in India or the section in the United States. The section forms the sampling unit and provides the means for selecting a sample of farms.

1.5 Simple Random Sampling

The simplest of the methods of probability sampling which provides estimates of the population characters and a measure of the reliability of the estimates made is the method of *simple random sampling*. In this method, usually called for brevity the method of random sampling, an equal probability of selection is assigned to each unit of the population at the first draw. The method implies an equal probability of selecting any unit from among the available units at subsequent draws. Thus, if the number of units in the population is N, the probability of selecting any unit at the first draw will be $1/N$, the probability of selecting any unit from among the available units at the second draw is $1/(N-1)$, and so on.

An important property of simple random sampling is that the probability of selecting a specified unit of the population at any given draw is equal to the probability of selecting it at the first draw. For, let

n denote the number of units to form the sample.

The probability that the specified unit is selected at the r-th draw is clearly the product of (1) the probability of the event that it is not selected in any of the previous $r-1$ draws; and (2) the probability of the event that it is selected at the r-th draw. The probability that it is not selected at the first draw is, by definition, $(N-1)/N$; that it is not selected at the second draw $(N-2)/(N-1)$, and so on. The probability of event (1) is, therefore,

$$\frac{N-1}{N} \cdot \frac{N-2}{N-1} \cdots \frac{N-r+1}{N-r+2}$$

The probability of the event (2) is clearly $1/(N-r+1)$. The product of the two is, therefore,

$$\frac{N-1}{N} \cdot \frac{N-2}{N-1} \cdots \frac{N-r+1}{N-r+2} \cdot \frac{1}{N-r+1} \text{ or } \frac{1}{N}$$

which is the probability of drawing the specified unit at the first draw.

Since the specified unit may be included in the sample at any of the n draws, it also follows that the probability that it is included in the sample is the sum of the probabilities that it is selected in the first draw, second draw, ..., n-th draw, and is, therefore, equal to n/N. Since this result is independent of the specified unit, it follows that every one of the units in the population has the same chance of being included in the sample under the procedure of simple random sampling. This, in fact, has sometimes been used as the definition of simple random sampling. However, this definition does not completely specify the procedure of simple random sampling, for, as will be clear in Chapter IX, there can be other procedures of sampling which do not give the same chance of selection at the first draw to each unit of the population and yet the probability that any specified unit is included in the sample is n/N.

The method of simple random sampling is also equivalent to giving an equal probability to each possible cluster of n units to form the sample of the population. The possible clusters of n are

$$\binom{N}{n}$$

Random sampling implies that every one of these possible clusters will have an equal probability, namely,

$$\frac{1}{\binom{N}{n}}$$

of being selected as the sample. Thus, if the population consists of 4 farms serially numbered 1, 2, 3 and 4, having 2, 3, 4 and 7 acres under corn respectively, then the possible clusters of 2 farms from this population will be the following six:

Serial Number of Cluster	Serial Number of Units in the Cluster	Values of the Units in the Cluster
1	1, 2	2, 3
2	1, 3	2, 4
3	1, 4	2, 7
4	2, 3	3, 4
5	2, 4	3, 7
6	3, 4	4, 7

Random sampling implies that every one of these 6 clusters of 2 each will have a chance of $\frac{1}{6}$ of being selected as the sample for our study. It is easy to establish this result.

The probability of selecting any one unit at the first draw is, by definition, $1/N$. Having selected one, the probability of selecting any one of the remaining units at the second draw is clearly $1/(N-1)$, and so on. The probability of selecting any given n units in succession in a specified order is thus

$$\frac{1}{N}\,\frac{1}{N-1}\cdots\frac{1}{N-n+1}$$

Since the order in which the units are selected is immaterial, the probability of any given n units to form the sample is thus given by

$$\frac{n!}{N\cdot(N-1)\ldots(N-n+1)} \text{ or } \frac{1}{\binom{N}{n}}$$

Every one of the $\binom{N}{n}$ possible clusters of n each has thus an equal probability of being selected under this method.

The word "random" refers to the method of selecting a sample rather than to the particular sample selected. Any possible sample can be a simple random sample, however unrepresentative it may appear, so long as it is obtained by following the rule of giving an equal chance to every one of the possible samples. Thus, a person may draw a sample of 13 cards from a well-shuffled pack and still find that all are of the same suit. The sample is obviously unrepresentative of different colours, but nevertheless must be considered to be a random sample by virtue of the method of selection employed.

1.6 Procedure of Selecting a Random Sample

A practical procedure of selecting a random sample is by using a table of random numbers, such as those published by Tippett (1927), a page from which is reproduced in the Appendix to this chapter. The procedure takes the form of (*a*) identifying N units in the population with the numbers 1 to N, or what is the same thing, preparing a list of units in the population and serially numbering them, (*b*) selecting *different* numbers from the table of random numbers, and (*c*) taking for the sample the *n* units whose numbers correspond to those drawn from the table of random numbers. The following examples will illustrate the procedure:

Example 1.1

Select a sample of 34 villages from a list of 338 villages.

Using the three-figure numbers given in columns 1 to 3, 4 to 6, etc., of the table given in the Appendix and rejecting numbers greater than 338 (and also the number 000), we have for the sample:

125, 326, 12, 237, 35, 251, 165, 131, 198,
33, 161, 209, 51, 52, 331, 218, 337, 263,
223, 241, 277, 42, 14, 303, 40, 99, 102,
173, 137, 321, 335, 155, 163, 81.

The procedure involves the rejection of a large number of random numbers, nearly two-thirds. A device commonly employed to avoid the rejection of such large numbers is to divide a random number by 338 and take the remainder as equivalent to the corresponding serial number between 1 to 337, the remainder zero corresponding to 338. It is, however, necessary to reject random numbers 677 to 999 and also 000 in adopting this procedure as otherwise villages with serial numbers 1 to 323 will get a larger chance of selection equal to 3/999 while those with serial numbers 324 to 338 will get a chance equal to 2/999. If we use this procedure and also the same three-figure random numbers as given in columns 1 to 3, 4 to 6, etc., we will obtain the sample of villages with serial numbers given below:

125, 206, 326, 193, 12, 237, 35, 251,
325, 338, 114, 231, 78, 112, 126, 330,
312, 165, 131, 198, 33, 161, 209, 51,
52, 331, 218, 337, 238, 323, 263, 90,
11, 223.

Example 1.2

The following procedure has been used for selecting a sample of fields for crop-cutting experiments on paddy in the surveys carried out by the Indian Council of Agricultural Research (1951).

"Against the name of each selected village are shown three random numbers smaller than the highest survey number* in the village. Select the survey numbers corresponding to given random numbers for experiments. If the selected survey number does not grow paddy, select the next higher paddy-growing survey number in its place."

Examine whether the above method will provide an equal chance of inclusion in the sample to all the paddy-growing survey numbers in the village, given the following:

1.	Name of village	Payagpur
2.	Total number of survey numbers	290
3.	Random numbers	18, 189, 239
4.	Paddy-growing survey numbers ..	49 to 88 and 189 to 290

Clearly, according to instructions, the survey numbers to be selected for crop-cutting experiments will be 49, 189 and 239. In selecting the first paddy-growing survey number for experiment, we thus give the survey number 49 a chance of 49/290 of being included in the sample, the survey number 189 a chance of 101/290, while to the remaining paddy-growing survey numbers a chance of 1/290 each. If paddy is grown in patches covering several survey numbers, as in the present example, the method will result in giving a larger chance to the border fields of being included in the sample.

Example 1.3

Nine villages in a certain administrative area contain 793, 170, 970, 657, 1721, 1603, 864, 383 and 826 fields respectively. Make a random selection of 6 fields, using the method of random sampling.

The total number of fields in all the 9 villages is 7987. The first step in the selection of a random sample of fields is to have these serially numbered from 1 to 7987, by taking successive cumulative totals:

793, 963, 1933, 2590, 4311, 5914, 6778, 7161, 7987,

the 793 fields in village 1 being given the serial numbers 1 through 793, the 170 fields in village 2 being given the serial

* Each field or separate piece of land in a village bears an official number which is termed the 'survey number'.

numbers 794 through 963, and so on. A reference to the four-digit random numbers in columns 9 to 12 will then give the following sample of fields with serial numbers 7358, 922, 4112, 3596, 633 and 3999. The corresponding fields will be No. 197 from village 9, No. 129 from village 2, No. 1522 and No. 1006 from village 5, No. 633 from village 1 and No. 1409 from village 5.

It will be noted that the selection has actually proceeded in two stages, selecting a village in the first instance with probability proportional to the number of fields in the village and choosing, on the basis of the random number already selected, a field in the selected village, villages being sampled with replacement. It must, however, be remembered that this equivalence between the one- and two-stage sampling holds good only when the number of second-stage units to be selected from a first-stage unit of sampling is limited to one.

Example 1.4

The following method is laid down for locating and marking a random plot of area 33′ × 33′ in a field selected for crop-cutting experiments in India (I.C.A.R., 1951).

"Stand facing North with the field in front of you and to your right. Measure the length and the breadth of the field in feet and deduct 33′ from each. Select a pair of random numbers less than or equal to the remainders so obtained to locate the corner of the plot. Fix a peg at this corner, tie a string to it and stretch it along the length of the field away from the South-West corner of the field. Measure 33′ along it by means of a tape and put the cross-staff at this point. Turn the string round the cross-staff and stretch it at right-angles away from the South-West corner of the field and measure 33′ along it. Proceed in this manner until you reach the starting point of the plot by checking the distance between the fourth and the first corner."

Examine whether the method will give an equal chance to all the unit areas in the field, it being given that the field is a rectangular field, measuring 120′ × 100′.

The method implies a division of the field into $(120–33)\times(100–33)$ plots, each measuring 33′ × 33′, which are not distinct but overlapping. The fundamental requirement that the population should be divisible into units which are distinct so that every unit area of the population belongs to one and only one sampling unit is thus not fulfilled, with the consequence that the central areas get a relatively greater chance of selection than those near the border.

1.7 Non-Random Methods of Sampling

Methods of sampling, which are not based on laws of chance but in which units of the population to be included in the sample are determined by the personal judgment of the enumerator, are called purposive or non-random methods. An example of this method, where personal judgment is introduced in the selection of a sample, is provided by the old official method in India of selecting fields for sample-harvesting for determining the average yield of a crop. Under this method, the experimenter was required to select fields which, in his judgment, had an average crop. It was found that the experimenter tended to select fields which were poorer than the average when the season was good and better than the average when the season was bad. The result was a tendency to over-estimate yields in bad years and to under-estimate them in good years. The quota method of sampling, so extensively used in the United States of America in opinion surveys, is another example of this method. Here quotas are set up for the different categories of the population to be included in the sample and the selection of units from each category is left to the personal discretion of the enumerator. The method is convenient to use in practice. Its cost is also low relative to that of the method of probability sampling. However, the sample does not provide any means of judging the reliability of the estimates based thereon. If we want to have unbiased estimates of the population character whose accuracy can be measured from the samples themselves, probability sampling alone should be used.

1.8 Non-Sampling Errors

The accuracy of a result is affected not only by sampling errors arising from chance variation in the selection of the sample but also by (*a*) lack of precision in reporting observations, (*b*) incomplete or faulty canvassing of a designated random sample, and (*c*) faulty methods of estimation. These errors, particularly those under (*a*) and (*b*), are usually grouped under the heading "non-sampling errors". Deming (1944, 1950) has listed the different sources of errors and biases arising from (*a*). These, in his words, are principally due to arbitrariness in definition and variable performance of the man. An eye-estimate of the crop provides an example of this source of errors. Eye-estimate is a form of measurement which cannot, in the very nature of things, give a unique result even when the same field is observed at different times by the same enumerator. The result will depend upon the personal judgment of the enumerator, no matter how well he is trained and consequently there will be variation from enumerator to enumerator observing the same field and in repeated observations by the same enumerator. A character like damage to a crop in the field from rust will similarly involve a certain amount of arbitrariness in definition and, therefore, give variable response. Even with factual characters like the area under the crop in a field, there is found to be marked variation in performance of the same enumerator measuring the acreage at different times or of different enumerators measuring the same field. In an inspection carried out by the statistical staff to test the reliability of the area records maintained by the patwaris (village officials) in 61 villages selected at random in the Lucknow District (India), about 20% of the reports were found to be in disagreement. Part of the discrepancies could, of course, be explained by carelessness or even dishonesty but most of them were due to differing descriptions of the same situation given by different agencies (Sukhatme and Kishen, 1951).

An example of faulty canvassing of a selected sample is reported by Kiser (1934) who selected a random sample of households for studying morbidity. The relative frequency distribution of the size of households included in the sample and as revealed by the Census is given in Table 1.1, which shows that the sample is considerably deficient in the frequency of households of size 2. Kiser attributed the deficiency to the failure on the part of the enumerators to re-visit missed households in which childless married women working away from homes are likely to predominate.

A similar bias attributed to the poor execution by the field force of the selected sample arose in a survey for estimating the yield

TABLE 1.1

Relative Frequency Distribution of the Size of the Households

Size of Household	Percentage Frequency in Sample	Percentage Frequency in Census
2	19·4	26·8
3	25·9	26·5
4	23·5	21·9
5	15·4	13·0
6	8·1	5·9
7	3·5	3·2
8	1·9	1·4
9 and over	2·2	1·3

of wheat carried out in Uttar Pradesh (India) in 1943–44. The harvesting had already commenced when the investigators went to conduct experiments in the selected fields, with the result that the fields harvested in the sample contained a larger proportion of irrigated fields than was present in the population. For, it is usually the unirrigated fields which mature earlier and which had, therefore, been harvested by the time the investigators commenced their experiments. The instruction to select the next available field for experiment would have only resulted in increasing the preponderance of irrigated fields in the sample.

Faulty methods of estimation can be illustrated with reference to a plan for estimating the average yield of a crop per acre. If, for example, an equal chance is given to every field under the crop to be included in the sample and the average yield per acre of larger fields is different from that of smaller ones, then the average yield per acre estimated from the simple arithmetic mean of the yields per acre of the fields in the sample may be markedly different from the average yield per acre of the total area under the crop.

The net effect of these discrepancies on the value of the estimate may not, however, always be serious, particularly in cases where errors occur in both directions and there is a reasonable chance of their cancelling one another. Errors, particularly those due to variability in reporting, introduce an additional component of variability which goes to inflate the estimate of the sampling error. It is frequently found that these errors do not cancel out and that the net effect is a bias due to the tendency to uniformly report a higher (or lower) figure than the true unknown value. It is, therefore, important to control these errors as far as practicable. Methods of measuring and controlling non-sampling errors will be discussed in Chapter X. In this chapter, we shall give examples which show that the magnitude of non-sampling errors can sometimes be much larger than what is commonly supposed, thus emphasising the need for reducing them as far as possible.

Example 1.5

This example is taken from an experiment conducted by the Indian Council of Agricultural Research at Poona (India) for evolving a method of obtaining a representative sample of fibres from a bulk of wool for the study of three wool characteristics, *viz.*, length, fineness and medullation. The experiment consisted of preparing well mixed lots, each weighing about ·6 gram, from a commercial mass of wool. Each lot was combed out and spread on a velvet board in a uniform thin layer and divided into three approximately equal sections by referring to a scale placed across the fibres. The first section was used for drawing a sample of 200 individual fibres by method (*a*) which consisted of drawing individual fibres with the help of random numbers by reading on the scale placed across the fibres. The second section was used for drawing a sample by method (*b*) which consisted of drawing bunches of approximately 50 fibres each from 4 random positions of the spread-out wool. The third section was used for drawing bunches of 100, but we will not refer to the results of this section here. The fibres in the samples, as also those left behind in each section, were then measured for length.

Table 1.2 gives the results of experiments on 24 lots measured by the same observer. It will be seen that the sample estimates based on individual fibres as sampling units exceed the corresponding population values in all the 24 lots, although the difference between the two has varied rather considerably from lot to lot. The results show that there is a consistent overestimation in the method of sampling individual fibres. The results of sampling by method (*b*), however, show that in 12 out of 24 lots, the sample estimate is larger than the corresponding population value, in 11 cases it is lower, and in the one remaining lot the two are equal, thus showing absence of bias in the method of sampling by bunches. It is clear that some conscious or unconscious tendency to select longer fibres of wool is introduced when method (*a*) of sampling is adopted. The procedure of random sampling implies identification of each one of the fibres in the population with the serial numbers 1 to N, and then selecting a sample of fibres with the help of random numbers. This, however, is an impracticable procedure to adopt in the sampling of wool. A practicable procedure is the one that was actually followed of spreading the lot in a thin layer across a velvet and selecting fibres from random positions with the help of a scale placed across it. The method, however, gives scope to the observer to select one fibre out of the several possible fibres in the neighbourhood of the random position. This scope, and with it the bias, is reduced when sampling is done by bunches.

Example 1.6

This is taken from an investigation carried out in Krishna District of Madras State (India) for comparing the efficiency of plots of different sizes in large-scale sample surveys for estimating the average yield of paddy (Sukhatme, 1947). 36 villages, distributed equally among the 6 subdivisions of the district, were selected for the investigation. In each selected village, 3 fields were selected at random out of all the paddy-growing fields in the village, and within each field the following plots were marked at random:

(*a*) A rectangle of 50×20 (links)^2 (area 435·6 sq.ft.), which is the plot size adopted in official crop-sampling work in Madras;

(*b*) Two circles of radius 3 ft. each (area 28·3 sq.ft.); and

(*c*) Two circles of radius 2 ft. each (area 12·6 sq.ft.).

Besides, the whole of the remaining field was harvested. The rectangular plot was marked with the help of tapes and pegs in the manner described in Example 1.4. The circular plots were marked with the help of a special apparatus devised for the

TABLE 1.2

Mean Fibre Length in Cms. in the Sample and in the Population in 24 Lots of Wool in Sampling with Methods (a) and (b)

Lots	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Method (*a*)												
Sample	4·60	7·59	5·74	4·62	4·35	9·29	5·65	8·25	8·04	5·33	7·60	8·17
Population	4·35	6·07	5·24	4·59	4·21	7·49	5·30	7·10	6·16	4·86	6·10	6·25
Method (*b*)												
Sample	4·56	5·75	5·92	5·34	4·12	8·37	5·93	7·24	5·23	4·95	6·78	5·76
Population	4·78	5·75	5·59	4·47	4·14	7·41	5·64	7·01	5·78	4·89	6·49	5·85

Lots	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)	(21)	(22)	(23)	(24)
Method (*a*)												
Sample	7·92	5·41	6·11	6·95	6·86	5·71	6·85	5·74	7·58	6·31	6·84	7·03
Population	6·10	5·00	5·44	6·06	6·28	5·24	5·77	5·61	6·51	5·17	6·33	5·81
Method (*b*)												
Sample	6·16	4·93	5·76	5·73	5·73	5·43	5·40	5·46	6·42	4·18	6·11	6·23
Population	6·24	5·06	5·40	6·06	5·81	4·92	5·59	5·51	6·34	4·55	5·66	5·67

purpose, consisting of a rotating peg, a steel tape and a plumb line. The peg was made of wood and was provided with an iron collar at one end and a point at the other. It was fixed at a point in the field located by means of a pair of random numbers. The steel tape was so fixed as to revolve fully round the centre of the top of the peg. As the tape was revolved, the crop was cut from below the level of the tape, thus making room for the tape to move further until the original starting point was reached. To avoid trampling, the point located by means of a pair of random numbers was not taken as the centre of the circle but was taken as a point of its circumference on a line parallel to the length of the field. On arriving at this point, the worker was asked to cut the crop from this point along the direction of the length until he reached a distance slightly exceeding the radius from this point. From the starting point the worker then measured exactly a distance equal to the radius along the direction of the length of the field and fixed the peg at this point. This was the centre of the circle. The field work was carried out by the local staff of the Department of Agriculture who had been given thorough training prior to the commencement of the investigation. The results of the investigation are reproduced in Table 1.3.

TABLE 1.3

Average Yield of Paddy in Lb./Acre for Plots of Different Sizes Paddy Survey in Krishna District (Madras)

Size and Shape of Plot		Area in Sq.ft.	No. of Plots	Average Yield in Lb./Acre	Standard Error	Percentage Over-Estimation
Whole field	..	..	108	1939·2	107·3	..
50×20 (links)^2	..	435·60	108	1954·1	105·0	0·8
3′ circle	..	28·27	216	2025·9	125·8	4·5
2′ circle	..	12·57	216	2113·2	129·1	9·0

It is seen that while the yield estimate from the official plot size of 50 links × 20 links is in close agreement with that from harvesting the whole field, those from small plots are considerable over-estimates.

The instructions for locating the starting point and for marking of plot were as objective as they possibly could be. Nevertheless, any one who has had experience of measuring the length of a field and walking from a given point in the field along the direction of its length will agree that the starting point of the plot and the direction along which it is to be laid in the field could at best be determined only approximately. Even if the same observer were to locate and mark the plot determined by a given pair of random numbers at different times in the same field, the plots may occupy different positions. The inclusion or exclusion of particular plants on the border of the plot in demarcating it will similarly depend upon the judgment of the experimenter. The area actually cut may also vary from the one intended to be cut due to unevenly sown crops and errors in measurement. If all these deviations could be ascribed to a random element, one would expect the errors to cancel out. The results given in Table 1.3, however, indicate that this is not the case. They show that small plots significantly over-estimate the yield, although the degree of over-estimation becomes smaller with larger plots. It is obvious that the overall influence of the various non-sampling errors relative to the produce harvested becomes smaller with the increase in plot size until, when the plot size is large enough, such as is used in official crop-sampling work, the bias becomes negligible.

The above examples will show the need for exercising care in preparing the design of a survey so that, as far as possible, biases are absent. Where it is not practicable to ensure absence of bias, one should at least satisfy oneself that the bias present, if any, in the sample estimate is so small as to be negligible in comparison with its standard error.

REFERENCES

1. Tippett, L. H. C. (1927) .. *Random Sampling Numbers*, Tracts for Computers, XV, Cambridge University Press.
2. I.C.A.R., New Delhi (1951) *Sample Surveys for the Estimation of Yield of Food Crops* (1944–49), Bulletin No. 72.
3. Deming, W. E. (1944) .. "On Errors in Surveys," *Amer. Sociol. Rev.*, **9**, 359–69.
4. ——— (1950) .. *Some Theory of Sampling*, John Wiley & Sons, Ltd., New York.
5. Sukhatme, P. V. and Kishen, K. (1951) "Assesment of the Accuracy of Patwaris' Area Records," *Agriculture and Animal Husbandry*, **1**, No. 9, 36–47.
6. Kiser, C. V. (1934) .. "Pitfalls in Sampling for Population Study," *Jour. Amer. Statist. Assoc.*, **29**, 250–56.
7. Sukhatme, P. V. (1947) .. "The Problem of Plot Size in Large-Scale Yield Surveys," *Jour. Amer. Statist. Assoc.*, **42**, 297–310.

APPENDIX

Table of Random Numbers

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
1	1	2	5	4	2	8	5	8	7	3	5	8	4	0	2	4	3	6	8	4	8	4	8	5	2	6	1	7	5	4	8	8
2	5	4	4	3	4	9	1	1	0	9	2	2	7	1	3	4	4	7	9	8	1	3	1	1	8	7	0	1	2	2	1	0
3	3	2	6	2	2	3	2	2	4	1	1	2	9	8	7	7	4	7	7	6	4	5	1	2	1	7	4	6	2	5	9	3
4	7	8	0	9	0	2	9	7	8	9	5	6	2	1	5	8	7	7	8	0	0	7	5	3	1	2	3	2	7	1	8	1
5	6	8	6	2	4	1	9	4	3	5	9	6	5	0	7	2	4	4	7	3	3	0	9	9	0	7	2	9	4	9	5	0
6	9	1	7	9	3	8	1	4	9	1	5	3	2	1	2	7	6	7	4	5	9	6	4	6	8	1	0	5	3	1	3	3
7	5	3	1	7	0	9	8	6	0	6	3	3	6	4	8	0	4	8	3	4	8	7	1	0	8	8	2	9	8	5	7	2
8	0	1	2	6	4	7	7	7	8	0	3	4	9	2	1	7	2	1	2	8	2	2	3	2	5	0	3	9	8	6	3	7
9	2	3	7	2	7	7	7	4	9	4	4	6	7	1	7	8	8	4	0	3	3	9	7	1	0	8	9	9	5	2	7	4
10	0	3	5	7	5	2	7	6	3	9	9	9	0	2	6	1	9	2	5	5	5	7	8	0	5	7	2	8	0	0	3	2
11	7	8	5	5	9	7	0	7	5	2	5	9	4	2	6	3	9	8	7	8	4	9	1	8	0	9	8	7	9	1	1	8
12	2	5	1	0	4	2	5	4	1	5	4	3	0	2	2	4	0	1	1	2	6	5	2	3	8	6	6	7	4	7	0	7
13	6	6	3	9	1	9	1	3	3	1	2	0	9	1	4	9	6	1	4	5	5	8	9	5	0	7	2	6	3	8	8	3
14	6	7	6	9	1	4	3	5	9	1	0	7	4	7	6	2	9	9	0	2	3	7	6	4	7	3	8	8	2	7	2	9
15	4	5	2	7	8	0	0	0	8	6	4	8	3	3	6	6	7	9	4	5	4	8	4	7	4	3	1	7	9	6	3	6
16	5	6	9	9	9	8	8	3	2	4	5	6	0	8	9	3	4	1	3	2	6	6	6	8	0	7	9	9	6	1	3	7
17	4	1	6	0	1	4	4	5	2	8	8	7	0	7	2	4	1	2	9	4	8	9	8	8	1	5	2	7	1	4	6	7
18	4	5	0	6	2	4	7	4	3	5	9	0	5	3	0	8	7	6	4	0	7	1	2	8	1	0	2	3	2	4	1	8
19	4	6	4	5	0	6	1	3	9	8	4	6	4	4	5	8	5	6	6	6	7	6	7	1	1	1	8	4	2	3	2	8
20	6	6	8	6	3	5	4	4	9	8	2	8	9	1	8	7	0	5	0	6	6	4	7	3	5	3	5	6	8	9	4	0
21	6	5	0	3	0	3	2	9	7	8	9	9	8	2	1	1	0	8	5	2	8	0	6	6	5	7	0	6	1	9	4	0
22	1	6	5	8	4	2	8	8	1	8	5	6	5	3	1	9	3	5	1	2	8	9	8	1	7	4	6	8	3	8	3	6
23	1	3	1	9	1	2	0	4	3	3	4	4	8	8	8	6	3	9	4	6	6	7	7	7	1	7	0	0	0	3	2	3
24	5	0	3	0	4	0	2	7	2	0	7	7	9	8	1	2	8	6	4	5	3	2	9	0	7	1	9	1	6	1	5	2
25	7	8	6	6	2	0	2	9	5	1	5	6	2	0	0	3	2	9	4	0	0	2	3	7	7	6	7	0	2	8	5	2

APPENDIX (*Contd.*)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
26	1	9	8	3	8	9	9	2	1	0	1	7	7	2	6	3	7	6	9	9	4	1	5	1	8	1	3	2	7	2	7	1
27	9	9	4	4	0	8	4	5	7	4	6	8	3	9	3	6	8	0	0	2	0	8	5	7	5	7	8	4	4	4	8	0
28	0	3	3	0	9	9	1	3	4	9	9	0	7	7	9	0	6	9	3	2	0	8	7	1	1	9	8	8	9	8	8	1
29	9	9	0	3	7	9	1	4	4	1	3	8	6	8	2	6	0	2	3	0	1	3	3	7	7	4	1	3	8	8	4	0
30	1	6	1	4	7	8	6	2	9	5	0	0	4	1	0	9	1	0	3	7	2	9	7	8	6	0	7	5	0	9	7	1
31	2	0	9	6	1	1	6	4	3	7	8	8	6	2	5	7	0	6	3	2	0	6	9	3	2	2	6	3	5	2	9	0
32	0	5	1	1	0	2	2	9	5	9	5	1	6	8	0	8	1	4	0	9	7	6	2	4	4	9	0	3	4	6	9	2
33	0	5	2	4	4	0	5	6	9	1	4	0	6	3	7	1	2	0	9	9	8	2	9	0	3	6	1	1	6	5	0	1
34	7	3	8	1	7	3	8	6	6	5	6	8	1	5	6	8	4	1	6	0	0	4	2	9	3	4	8	8	3	7	4	1
35	3	3	1	1	3	7	3	3	7	8	8	2	6	9	8	5	7	8	7	4	7	2	6	4	4	5	8	7	9	5	9	1
36	6	8	7	4	2	5	3	4	7	4	8	5	9	5	9	6	9	0	8	6	2	7	0	1	4	9	6	7	1	5	8	8
37	8	9	8	7	3	1	2	1	3	6	2	8	0	3	7	2	1	0	5	9	6	3	3	9	8	9	7	3	4	2	1	8
38	9	2	0	5	2	3	5	8	0	3	9	3	3	1	9	6	1	6	1	2	8	3	9	7	4	3	9	0	0	1	8	7
39	9	7	4	9	7	9	8	3	7	0	7	6	9	7	9	1	1	5	3	0	8	1	2	7	4	4	7	4	1	8	9	5
40	2	1	8	3	2	1	0	9	2	8	7	4	5	7	3	3	1	5	6	7	7	7	6	4	4	9	3	9	9	9	1	9
41	6	9	2	6	3	0	8	5	2	0	7	9	3	3	3	0	4	4	3	2	9	5	2	4	0	3	2	7	9	6	4	0
42	3	3	7	3	3	5	6	7	0	3	7	1	5	9	3	2	3	9	2	3	7	2	5	0	8	5	7	8	5	8	6	9
43	9	7	7	1	5	5	4	2	4	7	1	5	5	5	2	7	3	7	6	3	3	1	6	7	3	6	7	9	4	3	9	9
44	9	3	5	3	5	5	7	6	5	4	7	4	0	1	9	0	7	2	7	4	6	9	9	3	3	9	2	0	7	2	7	2
45	5	7	6	1	6	3	0	1	3	5	5	8	6	2	0	5	3	0	1	2	6	1	9	5	8	4	6	1	2	0	4	6
46	6	8	5	0	8	1	2	2	4	4	5	5	2	9	4	0	9	9	4	5	9	6	8	8	3	5	8	8	8	3	1	1
47	6	6	1	6	6	7	6	0	4	9	3	8	0	0	6	6	0	3	9	1	8	8	9	8	4	7	5	3	7	4	0	2
48	2	6	3	3	4	2	5	5	8	7	5	5	8	4	3	4	4	3	3	4	8	9	9	2	1	2	6	0	0	5	4	7
49	6	8	3	7	4	8	9	8	9	5	2	7	3	5	2	6	6	5	3	6	7	7	0	3	9	9	4	1	8	4	5	4
50	4	5	2	3	7	7	7	2	3	8	8	8	7	2	4	3	1	3	3	0	7	1	1	5	4	8	9	7	6	6	1	8

CHAPTER II

BASIC THEORY

A. SIMPLE RANDOM SAMPLING

2*a*.1 Introduction and Notation

Simple random sampling is by far the most common method of sampling in surveys. It is operationally convenient and simple in theory, as the name suggests. We shall first present the basic theory of simple random sampling and then go on to consider briefly the theory of sampling in which the probabilities of selection are unequal. We shall give the theory in its application to both quantitative (*i.e.*, measurable) and qualitative characters. We shall assume, unless otherwise mentioned, that the sampling units are drawn without replacement.

Let

N denote the number of sampling units in the population,

y the character under consideration,

y_i the value of the character for the i-th sampling unit of the population,

$\bar{y}_N$ the mean value of the character per unit of the population given by

$$\bar{y}_N = \frac{\sum_{i=1}^{N} y_i}{N} \tag{1}$$

$\mathbf{S}^2$ the mean square for the population given by

$$\begin{aligned}\mathbf{S}^2 &= \frac{\sum_{i=1}^{N} (y_i - \bar{y}_N)^2}{N-1} \\ &= \frac{\sum_{i=1}^{N} y_i^2 - N\bar{y}_N^2}{N-1}\end{aligned} \tag{2}$$

$V(y)$ the variance of a single observation in the population given by

$$V(y) = \frac{\sum_{i=1}^{N}(y_i - \bar{y}_N)^2}{N} \tag{3}$$

$$= \frac{N-1}{N}\, S^2 \tag{4}$$

n the size, *i.e.*, the number of sampling units in the sample,

$\bar{y}_n$ the sample mean given by

$$\bar{y}_n = \frac{\sum^{n} y_i}{n} \tag{5}$$

and

s^2 the sample mean square given by

$$s^2 = \frac{\sum^{n}(y_i - \bar{y}_n)^2}{n-1}$$

$$= \frac{\sum^{n} y_i^2 - n\bar{y}_n^2}{n-1} \tag{6}$$

where the summations extend over all the units in the sample.

2*a*.2 Unbiased Estimates of the Population Values

An estimate will vary from sample to sample, depending upon the units included in the sample. Thus, for the population mentioned in Section 1.5, the sample means will be seen to vary from 2·5 to 5·5 acres per farm. The sample mean square will similarly be found to vary from 0·5 to 12·5, as shown in Table 2.1. It will, however, be seen that the averages of the sample means and sample mean squares over the totality of samples are equal to the corresponding population values. Such sample values are called unbiased estimates of the population values. Algebraically, this is expressed as:

$$E(\bar{y}_n) = \frac{\sum^{\binom{N}{n}} \bar{y}_n}{\binom{N}{n}}$$

$$= \bar{y}_N \qquad (7)$$

$$E(s^2) = \frac{\sum'^{\binom{N}{n}} s^2}{\binom{N}{n}}$$

$$= \mathbf{S}^2 \qquad (8)$$

where the symbol E stands, as usual, for expectation. We write

$$\text{Est. } \bar{y}_N = \bar{y}_n \qquad (9)$$

and

$$\text{Est. } \mathbf{S}^2 = s^2 \qquad (10)$$

but sometimes when it is more convenient, we shall also use the circumflex notation to denote the estimate, as $\hat{\bar{y}}_N$ and $\hat{\mathbf{S}}^2$. It will be shown in the following sections that, when a sample is selected by the method of simple random sampling, an unbiased estimate of the population mean is given by the sample mean and an unbiased estimate of the population mean square by the sample mean square.

TABLE 2.1

Values of the Mean and the Mean Square in Different Samples of Two from the Population Mentioned in Section 1.5

Serial No. of the Sample	Values of Units in the Sample	$\bar{y}_n$	s^2	$\bar{y}_n - \bar{y}_N$	$(\bar{y}_n - \bar{y}_N)^2$	$s^2 - \mathbf{S}^2$	$(s^2 - \mathbf{S}^2)^2$
1	2, 3	2·5	0·5	−1·5	2·25	−25/6	625/36
2	2, 4	3·0	2·0	−1·0	1·00	−16/6	256/36
3	2, 7	4·5	12·5	0·5	0·25	47/6	2209/36
4	3, 4	3·5	0·5	−0·5	0·25	−25/6	625/36
5	3, 7	5·0	8·0	1·0	1·00	20/6	400/36
6	4, 7	5·5	4·5	1·5	2·25	− 1/6	1/36
Total		24·0	28·0	0	7·0	0	4116/36
Mean		4·0	$4\frac{2}{3}$	0	$1\frac{1}{6}$	0	$19\cdot0\dot{5}$

2*a*.3* Markoff's Theorem

In the sequel, unless otherwise stated, we shall consider only linear combinations of the values for the units in the sample as unbiased estimates of the population values. Among these, we shall call that unbiased linear estimate the best which has the minimum sampling variance. It can be shown by a corollary of Markoff's theorem (Neyman and David, 1938) on unbiased linear estimates that the best unbiased linear estimate of the population mean $\bar{y}_N$ is given by that value of $\hat{\bar{y}}_N$ for which $u = \sum_{i=1}^{n} (y_i - \hat{\bar{y}}_N)^2$ is minimum, if $y_1, y_2, \ldots, y_n$ are n observations on n variates having the same variance and the same covariance with means given by $E(y_i) = \bar{y}_{N.}$ It is easily shown that u is minimum when $\hat{\bar{y}}_N = \bar{y}_n$.

2*a*.4 Expected Value of the Sample Mean

We write

$$E(\bar{y}_n) = E\left\{\frac{1}{n}\sum^{n} y_i\right\} \tag{11}$$

where y_i stands for the value of the i-th unit of the population, and the summation is taken over all the n units in the sample. Numbering the units in the sample serially, as $1, 2, \ldots, r, \ldots, n$, we may write (11) as

$$E(\bar{y}_n) = E\left\{\frac{1}{n}\sum_{r=1}^{n} y_r'\right\} \tag{12}$$

where y_r' now stands for the value of the unit included in the sample at the r-th draw.

By a well-known theorem in probability, the expected value of a sum is the sum of the expected values. We, therefore, write

$$E(\bar{y}_n) = \frac{1}{n}\left\{E(y_1')+E(y_2')+\cdots+E(y_r')+\cdots+E(y_n')\right\} \tag{13}$$

[*] This section may be skipped over at the first reading without losing continuity of the text.

Now, by definition,

$$E(y_r') = \sum_{i=1}^{N} P_{i_r} y_i \qquad (14)$$

where P_{i_r} denotes the probability of drawing a specified unit y_i at the r-th draw. We have seen in Section 1.5 that, in simple random sampling, this probability is equal to $1/N$. It follows, therefore, that

$$\begin{aligned} E(y_r') &= \frac{1}{N} \sum_{i=1}^{N} y_i \\ &= \bar{y}_N \qquad (r = 1, 2, \ldots, n) \end{aligned} \qquad (15)$$

On substituting from (15) in (13), we get

$$E(\bar{y}_n) = \bar{y}_N \qquad (16)$$

An alternative and, in many ways, a more instructive approach is to write the sample mean in the form:

$$\frac{1}{n} \sum^{n} y_i = \frac{1}{n} \left\{ \sum_{i=1}^{N} a_i y_i \right\} \qquad (17)$$

where

$a_i = 1$ if y_i is in the sample,

and

$a_i = 0$ otherwise.

From (17), taking the expected value of both sides, we write

$$E\left\{ \frac{1}{n} \sum^{n} y_i \right\} = \frac{1}{n} \left[\sum_{i=1}^{N} \{E(a_i)\} y_i \right] \qquad (18)$$

Now, clearly,

$$\begin{aligned} E(a_i) &= 1.\{\text{Probability that } y_i \text{ is included in the sample}\} \\ &= \frac{n}{N} \end{aligned} \qquad (19)$$

in virtue of Section 1.5.

Hence, substituting n/N for $E\,(a_i)$ in (18), we reach the same result as (16).

2a.5 Expected Value of the Sample Mean Square

By analogy with (14), we have

$$E\,(y_r'^2) = \frac{1}{N} \sum_{i=1}^{N} y_i^2 \tag{20}$$

Adding and subtracting $\bar{y}_N^2$ from the right-hand side and using (2), we can write (20) in an alternative form as:

$$E\,(y_r'^2) = \bar{y}_N^2 + \left(1 - \frac{1}{N}\right) S^2 \tag{21}$$

It follows that

$$E\left\{\frac{1}{n} \sum^{n} y_i^2\right\} = E\left\{\frac{1}{n} \sum_{r=1}^{n} y_r'^2\right\}$$

$$= \bar{y}_N^2 + \left(1 - \frac{1}{N}\right) S^2 \tag{22}$$

Again, if y_r' and y_s' are used to denote the values of the units drawn at the r-th and s-th draws respectively, say y_i and y_j, we have, by definition,

$$E\,(y_r'y_s') = \sum_{i \neq j=1}^{N} P_{i_r} \cdot P_{j_s|i} \cdot y_i y_j \tag{23}$$

where $P_{j_s|i}$ denotes the probability of drawing y_j at the s-th draw, given y_i.

Now, from Section 1.5, we have

$$P_{i_r} = \frac{1}{N} \tag{24}$$

and, by an extension of the same result,

$$P_{j_s|i} = \frac{1}{N-1} \tag{25}$$

Hence, substituting for P_{i_r} and $P_{j_{s|i}}$ from (24) and (25) in (23), we get

$$E(y_r'y_s') = \frac{1}{N(N-1)} \sum_{i \neq j = 1}^{N} y_i y_j \tag{26}$$

It follows that

$$\frac{1}{n(n-1)} E\left\{\sum_{i \neq j}^{n} y_i y_j\right\} = \frac{1}{n(n-1)} E\left\{\sum_{r \neq s = 1}^{n} y_r' y_s'\right\}$$

$$= \frac{1}{N(N-1)} \sum_{i \neq j = 1}^{N} y_i y_j \tag{27}$$

The result can be alternatively established as follows:

We have

$$\frac{1}{n(n-1)} E\left\{\sum_{i \neq j}^{n} y_i y_j\right\} = \frac{1}{n(n-1)} \left[\sum_{i \neq j = 1}^{N} E(\alpha_i \alpha_j)\, y_i y_j\right] \tag{28}$$

where the summation $\sum\limits_{i \neq j}^{n}$ extends over the $n(n-1)$ product terms $y_i y_j$ in the sample, and $E(\alpha_i \alpha_j)$ denotes the probability of including y_i and y_j in the sample. Now,

$$E(\alpha_i \alpha_j) = E(\alpha_i) \cdot E(\alpha_j \mid \alpha_i) \tag{29}$$

where $E(\alpha_j \mid \alpha_i)$ denotes the probability of including y_j in a sample of $(n-1)$, after y_i is already drawn, from $(N-1)$.

Clearly, by the same argument by which we derived the value of $E(\alpha_i)$, we have

$$E(\alpha_j \mid \alpha_i) = \frac{n-1}{N-1} \tag{30}$$

It follows, therefore, that

$$E(\alpha_i \alpha_j) = \frac{n(n-1)}{N(N-1)} \tag{31}$$

Hence, on substituting from (31) in (28), we have

$$\frac{1}{n(n-1)} E\left\{\sum_{i \neq j}^{n} y_i y_j\right\} = \frac{1}{N(N-1)} \sum_{i \neq j = 1}^{N} y_i y_j \tag{32}$$

which can, alternatively, be expressed as:

$$\frac{1}{n(n-1)} E\left\{\sum_{i \neq j}^{n} y_i y_j\right\} = \frac{1}{N(N-1)}\left\{\left(\sum_{i=1}^{N} y_i\right)^2 - \sum_{i=1}^{N} y_i^2\right\}$$

$$= \frac{1}{N(N-1)}$$

$$\times \left\{N^2\bar{y}_N^2 - N\bar{y}_N^2 - (N-1)\mathbf{S}^2\right\}$$

$$= \bar{y}_N^2 - \frac{\mathbf{S}^2}{N} \tag{33}$$

or

$$E(y_r' y_s') = \bar{y}_N^2 - \frac{\mathbf{S}^2}{N} \tag{34}$$

Finally, using the results in (22) and (33), we have

$$E(\bar{y}_n^2) = E\left\{\frac{1}{n}\sum^{n} y_i\right\}^2$$

$$= \frac{1}{n^2} E\left\{\sum^{n} y_i^2 + \sum_{i \neq j}^{n} y_i y_j\right\}$$

$$= \frac{1}{n^2}\left\{E\sum^{n} y_i^2 + E\sum_{i \neq j}^{n} y_i y_j\right\}$$

$$= \frac{1}{n^2}\left[n\left\{\bar{y}_N^2 + \left(1 - \frac{1}{N}\right)\mathbf{S}^2\right\} + n(n-1)\left\{\bar{y}_N^2 - \frac{\mathbf{S}^2}{N}\right\}\right]$$

$$= \frac{1}{n^2}\left[n\bar{y}_N^2 + \left(n - \frac{n}{N}\right)\mathbf{S}^2 + n(n-1)\bar{y}_N^2\right.$$

$$\left. - \frac{n(n-1)}{N}\mathbf{S}^2\right]$$

$$= \bar{y}_N^2 + \left(\frac{1}{n} - \frac{1}{N}\right)\mathbf{S}^2 \tag{35}$$

Hence

$$E(s^2) = E\left\{\frac{1}{n-1}\left(\sum^{n} y_i^2 - n\bar{y}_n^2\right)\right\}$$

$$= \frac{1}{n-1}\left\{E\sum^{n} y_i^2 - nE(\bar{y}_n^2)\right\}$$

$$= \frac{1}{n-1}\left\{n\bar{y}_N{}^2 + \left(n - \frac{n}{N}\right)\mathrm{S}^2 - n\bar{y}_N{}^2 - \left(1 - \frac{n}{N}\right)\mathrm{S}^2\right\}$$

$$= \mathrm{S}^2 \tag{36}$$

showing that s^2 is an unbiased estimate of S^2.

2*a*.6 Sampling Variance, Standard Error and Mean Square Error

We have seen that a sample estimate differs from the population value by varying magnitudes in different samples. This difference between the sample estimate and the population value is called the *sampling error* of the estimate. An important requirement of a sampling method is that, in addition to giving an estimate of the population value, it should provide a measure of the sampling error in the estimate. Since the actual sampling error in an estimate cannot be known, we obtain a measure of the average magnitude over all possible samples of the sampling error in the estimate. A simple average of the actual sampling errors over all possible samples is, however, zero in the case of unbiased estimates, as seen from Table 2.1. An average of the sampling error without regard to sign provides one measure, called the *mean deviation*, but this is not in common use. The average magnitude of the squares of sampling errors over all possible samples is called the *sampling variance* of the estimate and its square-root is the measure most commonly used for defining the average sampling error. This measure of the average sampling error is called the *standard error*. In defining the standard error as above, we assume that the sample estimate provides an unbiased estimate of the population value. More generally, for a biased estimate of the population value, the sampling variance is defined as the arithmetic mean of the squares of the differences between the sample estimate and the expected value of the estimate over all samples, and its square root is called the standard error. The arithmetic mean of the squares of the differences between the sample estimate and the population value, in this case, is called the *mean square error*.

2*a*.7 Sampling Variance of the Mean

Let $V(\bar{y}_n)$ denote the sampling variance of the mean. Then, by definition, we have

$$V(\bar{y}_n) = E[\{\bar{y}_n - E(\bar{y}_n)\}^2]$$

$$= E(\bar{y}_n^2) - \{E(\bar{y}_n)\}^2 \tag{37}$$

Substituting from (16) and (35), we obtain

$$V(\bar{y}_n) = \left(\frac{1}{n} - \frac{1}{N}\right) S^2 \tag{38}$$

which can also be written as

$$V(\bar{y}_n) = \frac{N-n}{N} \cdot \frac{S^2}{n} \tag{39}$$

The reader may verify that the value of the sampling variance derived from this formula is the value actually obtained in Table 2.1 for a sample of size 2.

The factor $(N - n)/N$ in (39) is a correction for the finite size of the population and is called the finite population correction factor or simply the *finite multiplier*. When n is small as compared with N, the multiplier will approach unity and the sampling variance of the mean will approximate to that for the mean of a sample drawn from an infinite population.

Usually, the value of S^2 will not be known. Its estimate from the sample will, therefore, have to be used in calculating the sampling variance. Thus

$$\text{Est. } V(\bar{y}_n) = \frac{N-n}{N} \cdot \frac{s^2}{n} \tag{40}$$

and the estimate of the standard error is given by

$$\text{Est. } S.E. (\bar{y}_n) = \sqrt{\frac{N-n}{N}} \cdot \frac{s}{\sqrt{n}} \tag{41}$$

This estimate, however, as will be clear from Section 2*a*.11, has a slight negative bias.

It is instructive to derive the above result using the known result for the infinite population. Let μ_1 and μ_2 represent the mean and the variance for the infinite population, so that

$$E(y) = \mu_1$$

and

$$E\,(y-\mu_1)^2=\mu_2$$

We may write

$$(\bar{y}_n-\mu_1)=(\bar{y}_n-\bar{y}_N)+(\bar{y}_N-\mu_1)$$

Squaring both sides, we have

$$(\bar{y}_n-\mu_1)^2=(\bar{y}_n-\bar{y}_N)^2+(\bar{y}_N-\mu_1)^2+2\,(\bar{y}_n-\bar{y}_N)\,(\bar{y}_N-\mu_1)$$

If we now regard the finite population N itself as a random sample from the infinite population and, for any given N, choose n at random out of N, then the sample of n also becomes a random sample from the infinite population. Taking expectations of both sides in two stages, first for fixed $y_1, y_2, \ldots, y_N$ and then over all possible samples of N from the infinite population, we write

$$\begin{aligned} E\,[E\,\{(\bar{y}_n-\mu_1)^2 \mid y_1, y_2, \ldots, y_N\}] = {} & E\,\{V\,(\bar{y}_n) \mid y_1, y_2, \ldots, y_N\} \\ & + E\,(\bar{y}_N-\mu_1)^2 \\ & + 2E\,(\bar{y}_N-\mu_1)\,E\,\{(\bar{y}_n-\bar{y}_N) \mid y_1, \ldots, y_N\} \end{aligned}$$

But we have already seen in (16) that $E\,(\bar{y}_n)=\bar{y}_N$, and so $E\,\{(\bar{y}_n-\bar{y}_N) \mid y_1, \ldots, y_N\}=0$. The last term in the above expression is therefore zero. Hence

$$\frac{\mu_2}{n}=E\,\{V\,(\bar{y}_n) \mid y_1, \ldots, y_N\}+\frac{\mu_2}{N}$$

or

$$E\,\{V\,(\bar{y}_n) \mid y_1, \ldots, y_N\}=\left(\frac{1}{n}-\frac{1}{N}\right)\mu_2$$

It follows that

$$V\,\{(\bar{y}_n) \mid y_1, \ldots, y_N\}=\left(\frac{1}{n}-\frac{1}{N}\right)\hat{\mu}_2$$

But the estimate of μ_2 in a sample of N is provided by $\mathbf{S}^2$. Hence

$$V\,\{(\bar{y}_n) \mid y_1, \ldots, y_N\}=\left(\frac{1}{n}-\frac{1}{N}\right)\mathbf{S}^2$$

2*a*.8* Partitional Notation

An expression for the sampling variance of s^2 can be similarly derived. The calculation, however, involves much heavier algebra than in the case of the mean. The calculation of the sampling variance of the higher order moments is even more laborious. Their derivation is greatly facilitated by the use of partitional notation. We shall therefore digress to introduce this notation here and illustrate its application to derive the sampling variance of s^2.

A partition of a number w is a collection of positive integers 1 to 9 whose sum is equal to w. The integers are written in a descending order of magnitude and enclosed in brackets ( ). A partition which has ρ parts is called a ρ-part partition and the number w is called the weight of the partition. A repetition of the same part is indicated by exponents.

Examples:

(i) (2) and (1^2) are 1-part and 2-part partitions of 2 respectively.

(ii) (3), (2 1) and (1^3) are 1-part, 2-part and 3-part partitions of 3 respectively.

(iii) In general, $(p_1^{\pi_1} p_2^{\pi_2} \ldots p_h^{\pi_h})$ is a ρ-part partition of the number w, where p_i is repeated π_i times ($i = 1, 2, \ldots, h$), $\rho = \sum_{i=1}^{h} \pi_i$ and $w = \sum_{i=1}^{h} p_i\pi_i$.

The functions of observations we require are the monomial symmetric functions, called the g-functions in the notation of this book, and are written as $g\,(p_1^{\pi_1} p_2^{\pi_2} \ldots p_h^{\pi_h})$. Thus $\sum_{i=1}^{n} y_i^3$ will be denoted by $g\,(3)$, $\sum_{i\neq j=1}^{n} y_i^2 y_j$ by g (21), and $\sum_{i\neq j\neq k=1}^{n} y_i\, y_j\, y_k$ by $3!\, g\,(1^3)$. In general, the monomial symmetric function

$$\sum_{i_1\neq i_2\ldots\neq i_\rho=1}^{n} (y_{i_1}^{p_1} y_{i_2}^{p_1} \ldots y_{i_{\pi_1}}^{p_1} y_{i_{\pi_1+1}}^{p_2} \ldots y_{i_{\pi_1+\pi_2}}^{p_2} y_{i_{\pi_1+\pi_2+1}}^{p_3} \ldots y_{i_{\pi_1+\pi_2+\cdots\pi_h}}^{p_h})$$

will be denoted by

$$\pi_1!\, \pi_2! \ldots \pi_h!\ g\,(p_1^{\pi_1} p_2^{\pi_2} \ldots p_h^{\pi_h}) \qquad (42)$$

The special case of the g-function involving one-part partitions and thus denoting sums of powers of observations is called the s-function, and is denoted by replacing g by s, as $s\ (p)$, or sometimes by putting p as a suffix, *e.g.*, s_p. A product of two or more s-functions like $s_{p_1}, s_{p_2}, \ldots, s_{p_h}$ is written as $s\ (p_1\ p_2 \ldots p_h)$, where $(p_1\ p_2 \ldots p_h)$ is an h-part partition of the number $p_1 + p_2 + \ldots + p_h$.

Examples:

$$s\ (3) = \sum_{i=1}^{n} y_i^3$$

$$s\ (21) = s_2 s_1 = \left(\sum_{i=1}^{n} y_i^2\right)\left(\sum_{i=1}^{n} y_i\right)$$

$$s\ (1^3) = \{s\ (1)\}^3 = \left\{\sum_{i=1}^{n} y_i\right\}^3$$

In general,

$$s\ (p_1^{\pi_1}\ p_2^{\pi_2} \ldots p_h^{\pi_h}) = \{s\ (p_1)\}^{\pi_1} \{s\ (p_2)\}^{\pi_2} \ldots \{s\ (p_h)\}^{\pi_h}$$

$$= \left\{\sum_{i=1}^{n} y_i^{p_1}\right\}^{\pi_1} \left\{\sum_{i=1}^{n} y_i^{p_2}\right\}^{\pi_2} \cdots \left\{\sum_{i=1}^{n} y_i^{p_h}\right\}^{\pi_h}$$

We shall use capital symbols G and S to denote functions of observations for the population; small symbols, g and s, denoting as above the corresponding functions for the sample.

The g-functions can be expressed as linear functions of the products of s-functions, and *vice versa*, by means of the following identity:

$$-\sum_{i} \log (1 - y_i \alpha) \equiv s_1 \alpha + s_2 \frac{\alpha^2}{2} + s_3 \frac{\alpha^3}{3} + \ldots$$

where α is any constant such that

$$|\,\alpha\,| < \frac{1}{\max y_i}$$

giving us

$$g\ (p_1^{\pi_1}\ p_2^{\pi_2} \ldots) = \Sigma\, g_s\ (P,\ Q)\ s\ (q_1^{\chi_1}\ q_2^{\chi_2} \ldots) \qquad (43)$$

and

$$s\ (p_1^{\pi_1}\ p_2^{\pi_2} \ldots) = \Sigma\, s_g\ (P,\ Q)\ g\ (q_1^{\chi_1}\ q_2^{\chi_2} \ldots) \qquad (44)$$

where P and Q stand for the partitions $(p_1^{\pi_1} p_2^{\pi_2} \ldots)$ and $(q_1^{\chi_1} q_2^{\chi_2} \ldots)$ of the same number w given by

$$w = \Sigma\, p_i \pi_i = \Sigma\, q_i \chi_i \tag{45}$$

$\Sigma \pi_i$ and $\Sigma \chi_i$ denote the numbers of parts in the partitions P and Q respectively, $g_s\,(P, Q)$ denotes the coefficient of $s\,(q_1^{\chi_1} q_2^{\chi_2} \ldots)$ in the expansion of $g\,(p_1^{\pi_1} p_2^{\pi_2} \ldots)$, $s_g\,(P, Q)$ denotes the coefficient of $g\,(q_1^{\chi_1} q_2^{\chi_2} \ldots)$ in the expansion of $s\,(p_1^{\pi_1} p_2^{\pi_2} \ldots)$, and (43) and (44) are summed over all the different partitions Q of w, including P.

Examples:

$$g\,(1^2) = -\tfrac{1}{2}s_2 + \tfrac{1}{2}s_1^{\,2}$$

$$g\,(21^2) = s_4 - s_3 s_1 - \tfrac{1}{2}s_2^{\,2} + \tfrac{1}{2}s_2 s_1^{\,2}$$

$$s\,(2^2) = g\,(4) + 2g\,(2^2)$$

$$s\,(321) = g\,(6) + g\,(51) + g\,(42) + 2g\,(3^2) + g\,(321)$$

Values of the coefficients $s_g\,(P, Q)/\chi_1!\,\chi_2!\ldots$and $\pi_1!\,\pi_2!\ldots g_s\,(P, Q)$ for weights less than or equal to 8 have been tabulated (Sukhatme, 1938) and reproduced in the Appendix to this chapter. Expansions of g-functions in terms of s-functions, and *vice versa*, can be readily written down with the help of these tables.

2a.9* A Theorem on Expected Values of Symmetric Functions

We shall now state and prove the following theorem relating to the expected value of a g-function in samples selected by the method of simple random sampling from a finite population.

Theorem.—The expected value of a monomial symmetric function of the sample observations is e_ρ times the corresponding function for the population where

$$e_\rho = \frac{n\,(n-1)\ldots(n-\rho+1)}{N\,(N-1)\ldots(N-\rho+1)} \tag{46}$$

i.e.,

$$E\,\{g\,(p_1^{\pi_1} p_2^{\pi_2} \ldots\ p_h^{\pi_h})\} = e_\rho\, G\,(p_1^{\pi_1} p_2^{\pi_2} \ldots\ p_h^{\pi_h}) \tag{47}$$

Proof :

$$E\{g(p_1^{\pi_1} p_2^{\pi_2} \ldots p_h^{\pi_h})\}$$

$$= \frac{1}{\pi_1!\,\pi_2! \ldots \pi_h!}\, E\left\{\sum_{i_1 \neq i_2 \cdots \neq i_\rho = 1}^{n} (y_{i_1}^{p_1} y_{i_2}^{p_1} \ldots y_{i_{\pi_1}}^{p_1} y_{i_{\pi_1+1}}^{p_2} \ldots y_{i_{\pi_1+\pi_2+\cdots\pi_h}}^{p_h})\right\}$$

$$= \frac{1}{\pi_1!\,\pi_2! \ldots \pi_h!}\, E\left\{\sum_{i_1 \neq i_2 \cdots \neq i_\rho = 1}^{N} \alpha\, (y_{i_1}^{p_1} y_{i_2}^{p_1} \ldots y_{i_{\pi_1}}^{p_1} y_{i_{\pi_1+1}}^{p_2} \ldots y_{i_{\pi_1+\pi_2+\cdots\pi_h}}^{p_h})\right\}$$

where α is a simplified notation for the product

$$\alpha_{i_1}\, \alpha_{i_2} \ldots \alpha_{i_{\pi_1+\pi_2+\cdots\pi_h}}$$

$\alpha = 1$, if $y_{i_1}, y_{i_2}, \ldots, y_{i_{\pi_1+\pi_2+\cdots\pi_h}}$ are in the sample,

and

$\alpha = 0$ otherwise.

Hence

$$E\{g(p_1^{\pi_1} p_2^{\pi_2} \ldots p_h^{\pi_h})\}$$

$$= \frac{1}{\pi_1! \ldots \pi_h!} \sum_{i_1 \neq i_2 \cdots \neq i_\rho = 1}^{N} E(\alpha)\, (y_{i_1}^{p_1}\, y_{i_2}^{p_1} \ldots y_{i_{\pi_1}}^{p_1}\, y_{i_{\pi_1+1}}^{p_2} \ldots y_{i_{\pi_1+\pi_2+\cdots\pi_h}}^{p_h})$$

But

$$E(\alpha) = 1 \cdot \text{Probability } (\alpha = 1)$$

$$= \frac{n(n-1) \ldots (n-\rho+1)}{N(N-1) \ldots (N-\rho+1)}$$

$$= e_\rho$$

Hence

$$E\{g(p_1^{\pi_1} p_2^{\pi_2} \ldots p_h^{\pi_h})\} = E(\alpha)\, G(p_1^{\pi_1} p_2^{\pi_2} \ldots p_h^{\pi_h})$$
$$= e_\rho\, G(p_1^{\pi_1} p_2^{\pi_2} \ldots r_h^{\pi_h})$$

2*a*.10* Sampling Variance of s²

By definition,

$$V(s^2) = E\{(s^2)^2\} - \{E(s^2)\}^2$$

$$= E\{(s^2)^2\} - S^4 \tag{48}$$

Now

$$
\begin{aligned}
(n-1)^2 (s^2)^2 &= \left\{ \sum^{n} (y_i^2) - n\bar{y}_n^{\,2} \right\}^2 \\
&= \left(s_2 - \frac{s_1^{\,2}}{n} \right)^2 \\
&= \left(s_2^{\,2} - 2\,\frac{s_2 s_1^{\,2}}{n} + \frac{s_1^{\,4}}{n^2} \right) \qquad (49)
\end{aligned}
$$

A reference to $s_g\,(P, Q)/\chi_1!\,\chi_2!\dots$ table of weight 4 gives

$$
\begin{aligned}
(n-1)^2 (s^2)^2 = \{g\,(4) + 2g\,(2^2)\} - \frac{2}{n}\{g\,(4) + 2g\,(31) + 2g\,(2^2) \\
+ 2g\,(21^2)\} + \frac{1}{n^2}\{g\,(4) + 4g\,(31) + 6g\,(2^2) + 12g\,(21^2) \\
+ 24g\,(1^4)\}
\end{aligned}
$$

whence, taking expectations of both sides and using the theorem of the previous section, we have

$$
\begin{aligned}
(n-1)^2\,E\,\{(s^2)^2\} = \frac{(n-1)^2}{n^2}\,e_1 G\,(4) - \frac{4\,(n-1)}{n^2}\,e_2 G\,(31) \\
+ \frac{2}{n^2}\,(n^2 - 2n + 3)\,e_2 G\,(2^2) - \frac{4}{n^2}\,(n-3)\,e_3 G\,(21^2) \\
+ \frac{24}{n^2}\,e_4 G\,(1^4) \qquad (50)
\end{aligned}
$$

It is customary to express the expected values in terms of the population moments about the mean. This is done with the help of the table for $\pi_1!\,\pi_2!\dots G_S\;(P, Q)$ in the Appendix. We may, for simplicity, assume without loss of generality that the population mean, or S_1, is zero.

Substituting for the G functions the S functions, by using the aforesaid table of weight 4, we obtain

$$
\begin{aligned}
(n-1)^2\,E\,\{(s^2)^2\} = \frac{(n-1)^2}{n^2}\,e_1 S_4 + \frac{4\,(n-1)}{n^2}\,e_2 S_4 \\
+ \frac{n^2 - 2n + 3}{n^2}\,e_2\;(-S_4 + S_2^{\,2}) - \frac{2\,(n-3)}{n^2}\,e_3\,(2S_4 - S_2^{\,2}) \\
+ \frac{1}{n^2}\,e_4\,(-6S_4 + 3S_2^{\,2}) \qquad (51)
\end{aligned}
$$

Collecting terms and remembering that $S_4 = N\mu_4$, $S_2 = N\mu_2$ and $S^2 = N\mu_2/(N-1)$, we have

$$V(s^2) = \frac{n^2}{(n-1)^2}\left\{N\mu_4\left(\frac{e_1-e_2}{n^2} - \frac{2(e_1-3e_2+2e_3)}{n^3} + \frac{e_1-7e_2+12e_3-6e_4}{n^4}\right) + N^2\mu_2^2\left(\frac{e_2}{n^2} - \frac{2(e_2-e_3)}{n^3} + \frac{3(e_2-2e_3+e_4)}{n^4}\right)\right\} - \frac{N^2}{(N-1)^2}\mu_2^2 \tag{52}$$

It can be verified that the value 19·05 for the sampling variance of s^2 obtained in Table 2.1 agrees with that derived from the above formula.

The corresponding formula when N is infinite is readily obtained as the limiting case as $N \rightarrow \infty$. In this case, we have

$$N^i e_j = 0 \text{ for } i < j$$

and

$$N^j e_j = n(n-1)\ldots(n-j+1)$$

Hence

$$V(s^2) = \frac{\mu_4 - \mu_2^2}{n} + \frac{2}{n(n-1)}\mu_2^2 \tag{53}$$

Using the Pearsonian notation for departure from normality, this can be written as

$$V(s^2) = S^4\left\{\frac{\beta_2 - 1}{n} + \frac{2}{n(n-1)}\right\} \tag{54}$$

where

$$\beta_2 = \frac{\mu_4}{S^4}$$

For the normal population $\beta_2 = 3$, so that

$$V(s^2) = \frac{2}{n-1} S^4 \tag{55}$$

and

$$S.E.(s^2) = \sqrt{\frac{2}{n-1}}\, S^2 \tag{56}$$

Expected values of higher order sample moments and their products have been worked out by adopting the above procedure and tabulated for ready reference by Sukhatme (1944).

2*a*.11 Expected Value and Sampling Variance of *s*

We have seen that the sample mean square s^2 provides an unbiased estimate of S^2; and that, in samples from large populations, its sampling variance is given by

$$V(s^2) = \frac{\mu_4 - \mu_2^2}{n} + \frac{2}{n(n-1)}\mu_2^2$$

where μ_2 and μ_4 denote the second and the fourth moments of the population. However, sometimes, we also need to know the behaviour of s, the standard deviation. This is obtained as follows:

Let

$$s^2 = S^2 + \epsilon$$

where

$$E(\epsilon) = 0$$

and

$$E(\epsilon^2) = V(s^2)$$

We may write

$$s = (S^2 + \epsilon)^{\frac{1}{2}}$$

$$= S\left(1 + \frac{\epsilon}{S^2}\right)^{\frac{1}{2}}$$

Since ϵ will be small as compared with S^2 with a probability approaching 1 as n becomes large, we may expand the right-hand side as a series, neglecting powers of ϵ higher than the second. We then have

$$s = S\left\{1 + \tfrac{1}{2}\cdot\frac{\epsilon}{S^2} + \frac{\frac{1}{2}(-\frac{1}{2})}{1\cdot 2}\cdot\left(\frac{\epsilon}{S^2}\right)^2 + \cdots\right\}$$

Taking expectations of both sides, we obtain

$$E(s) = S\left\{1 - \frac{1}{8}\,\frac{V(s^2)}{S^4}\right\} \quad \ldots \quad (57)$$

The result shows that s will under-estimate S although, if n is large, the bias will be negligible.

Turning now to the evaluation of the sampling variance of s, we have

$$\begin{aligned} V(s) &= E\{s - E(s)\}^2 \\ &= E(s^2) - \{E(s)\}^2 \\ &= \mathrm{S}^2 - \mathrm{S}^2\left\{1 - \tfrac{1}{8} \cdot \frac{V(s^2)}{\mathrm{S}^4}\right\}^2 \\ &= \mathrm{S}^2\left\{1 - 1 + \frac{2V(s^2)}{8\mathrm{S}^4} \cdots\right\} \\ &\cong \frac{V(s^2)}{4\mathrm{S}^2} \qquad (58) \end{aligned}$$

where $V(s^2)$ denotes the variance of s^2 to terms up to $1/n$ only.

2*a*.12 Confidence Limits

The standard error gives an idea of the frequency with which errors (differences between the sample estimate and the population value) of a given magnitude may be expected to occur if repeated random samples of the same size are drawn from the population. Usually errors smaller than the standard error will occur with a frequency of about 68%, and those smaller than twice the magnitude of the standard error will occur with a frequency of about 95%, provided the estimate is approximately normally distributed. In general, if the sample size is not too small and N is large and if the estimate under consideration is a linear unbiased estimate of the population value, then the frequency with which errors will exceed a fixed multiple of the standard error of the estimate is approximately equal to the frequency as determined by the normal law. Consequently, from a knowledge of the standard error of the estimate and with the help of the normal probability integral tables, we are in a position to locate the actual unknown population value within certain limits with a known relative frequency. To take the example of estimating the population mean, we know that the mean of a random sample will be approximately normally distributed if the size of the sample is not too small and if the population from which it is drawn is not very different from the normal. We may, therefore, expect that

$$|\bar{y}_n - \bar{y}_N| \leqslant \sqrt{\frac{N-n}{Nn}}\,\mathrm{S} \tag{59}$$

on an average in 68 out of 100 occasions, and

$$|\bar{y}_n - \bar{y}_N| \leqslant 2\sqrt{\frac{N-n}{Nn}}\,\mathrm{S} \tag{60}$$

on an average with a frequency of about 95 out of 100. In general, we can expect the inequality

$$\bar{y}_n - t_{(\alpha,\,\infty)}\sqrt{\frac{N-n}{Nn}}\,\mathrm{S} \leqslant \bar{y}_N \leqslant \bar{y}_n + t_{(\alpha,\,\infty)}\sqrt{\frac{N-n}{Nn}}\,\mathrm{S} \tag{61}$$

where $t_{(\alpha,\,\infty)}$ is the value of the normal variate corresponding to the value $1 - \alpha/2$ of the normal probability integral, to hold on an average with a probability $1 - \alpha$. The two limits, on either side of the population mean in (61), are called the *confidence limits* and the interval between them the *confidence interval.* The probability with which the inequality holds, *viz.*, $1 - \alpha$, is termed the *confidence coefficient.*

It should be noted that the confidence limits may vary from sample to sample. Thus the confidence limits for the six different samples mentioned in Section 1.5 at the 68% and 95% confidence coefficients work out as shown in cols. 2 and 4 of Table 2.2.

TABLE 2.2

Confidence Limits for Different Samples Mentioned in Table 2.1

Sample No.	Confidence Limits ($1 - \alpha = \cdot 68$)		Confidence Limits ($1 - \alpha = \cdot 95$)	
	Based on S	Based on s	Based on S	Based on s
(1)	(2)	(3)	(4)	(5)
1	1·4, 3·6	1·8, 3·2	0·3, 4·7	− 2·0, 7·0
2	1·9, 4·1	1·7, 4·3	0·8, 5·2	− 6·0, 12·0
3	3·4, 5·6	1·2, 7·8	2·3, 6·7	−18·0, 27·0
4	2·4, 4·6	2·8, 4·2	1·3, 5·7	− 1·0, 8·0
5	3·9, 6·1	2·4, 7·6	2·8, 7·2	−13·0, 23·0
6	4·4, 6·6	3·5, 7·5	3·3, 7·7	− 8·0, 19·0

It will be observed that in four out of six cases, the population mean is contained within the confidence limits given in col. 2, while in all the six cases it is contained within the limits shown in col. 4, as is to be expected. The result is of course fortuitous in view of the small size of the population but it serves to demonstrate the meaning of the inequalities above.

When S^2 is not known, we use its estimate s^2 obtained from the sample. The statement in (61) with S^2 replaced by its estimate s^2 will, however, no longer be exact. To obtain the confidence limits in this case, we make use of the result that $(\bar{y}_n - \bar{y}_N)/S.E.\,(\bar{y}_n)$ is distributed as Student's t with $(n-1)$ degrees of freedom when n is not too small and the original distribution is not far removed from the normal. If we denote by $t_{(\alpha,\, n-1)}$ the value of t corresponding to the level of significance α for $(n-1)$ degrees of freedom, it follows that we may expect the inequality

$$\frac{|\bar{y}_n - \bar{y}_N|}{\sqrt{\frac{N-n}{Nn}}\, s} \leqslant t_{(\alpha,\, n-1)} \tag{62}$$

to hold on the average with probability $(1-\alpha)$. The $(1-\alpha)$ confidence limits when the size of the sample is not too small and the population from which it is drawn is not very different from the normal are, therefore, given approximately as

$$\bar{y}_n - t_{(\alpha,\, n-1)} \sqrt{\frac{N-n}{Nn}}\, s \leqslant \bar{y}_N \leqslant \bar{y}_n + t_{(\alpha,\, n-1)} \sqrt{\frac{N-n}{Nn}}\, s \tag{63}$$

For the six samples in Section 1.5 and for the same confidence coefficients as given above, these confidence limits based on s^2 are given in cols. 3 and 5 of Table 2.2. The values of $t_{(\cdot 32,1)}$ and $t_{(\cdot 05,\, 1)}$ have been interpolated from the t-table, being 1·85 and 12·7 respectively (Fisher and Yates, 1938).

2*a*.13 Size of Sample for Specified Precision

Almost the first question which a statistician is called upon to answer in planning a sample survey is about the size of the sample required for estimating the population value with a specified precision. The precision is usually specified in terms of the margin of error permissible in the estimate and the coefficient of confidence with which one wants to make sure that the estimate is within the permissible margin of error. Thus, if the error permissible in the estimate of the population value of the mean is, say, $\epsilon\bar{y}_N$ and the degree of assurance desired is $1 - \alpha$, then clearly we need to know the size of the sample so that

$$P\{|\bar{y}_n - \bar{y}_N| \geqslant \epsilon\bar{y}_N\} = \alpha \tag{64}$$

Hence, from (61), we have

$$n = \frac{\dfrac{t^2_{(\alpha,\infty)}}{\epsilon^2}\dfrac{\mathbf{S}^2}{\bar{y}_N^2}}{1 + \dfrac{1}{N}\dfrac{t^2_{(\alpha,\infty)}}{\epsilon^2}\dfrac{\mathbf{S}^2}{\bar{y}_N^2}} \tag{65}$$

The determination of the size of sample from (65) presumes the knowledge of the coefficient of variation for the population. This can only be roughly estimated. Consequently, (65) can give only a rough idea of the size of the sample required for estimating the population mean with a specified precision. We can, however, improve upon the predicted value of n as follows:

Although the size of the sample is determined from (65), the confidence limits after the survey is completed are obtained from (63). In other words, n could be more precisely evaluated from

$$n = \frac{\dfrac{t^2_{(\alpha,n-1)}}{\epsilon^2}\dfrac{s^2}{\bar{y}_N^2}}{1 + \dfrac{1}{N}\dfrac{t^2_{(\alpha,n-1)}}{\epsilon^2}\dfrac{s^2}{\bar{y}_N^2}} \tag{66}$$

had $t_{(\alpha,\, n-1)}$ been known, which it is not, as it itself depends upon n. As a result n is underestimated since $t_{(\alpha,\infty)}$ is less than $t_{(\alpha,\, n-1)}$. The obvious correction which suggests itself is to increase the value of n in the ratio $t^2_{(\alpha,n'-1)}/t^2_{(\alpha,\,\infty)}$, where n' is evaluated from (65), but the correction is not likely to be important unless n is small.

The calculation of n from (65) also assumes knowledge of $\mathbf{S}$ when the error, $\epsilon\bar{y}_N$, permissible in the estimate of the population value of the mean is given, although the confidence limits after the completion of the survey are calculated from (63) which makes use of s.

An allowance for this inaccuracy can be made by making use of the idea, originally due to Neyman (1934), of selecting a preliminary sample for improving the sampling design of the survey (Sukhatme, 1935). Let n_1 be the size of the preliminary sample and s_1^2 denote the estimate of $\mathbf{S}^2$ obtained therefrom. Then the additional sample required for estimating the population value with the desired accuracy, assuming N to be large and $\epsilon\bar{y}_N$ to be given, will be $n - n_1$, where

$$n = \frac{t^2_{(\alpha,\, n_1-1)}}{\epsilon^2} \frac{s_1^2}{\bar{y}_N^2} \tag{67}$$

It has been shown that n so estimated satisfies the statement in (64) and, on the average, gives a more accurate confidence interval than when $\mathbf{S}$ is unknown, but further discussion of the problem is beyond the scope of this book.

2a.14 Hyper-Geometric Distribution—Two Classes

We shall now consider the theory of simple random sampling as applied to qualitative characters. Consider, first, a situation in which the sampling units in the population are divided into two mutually exclusive classes, class 1 consisting of units possessing the attribute under consideration, and class 2 consisting of those not possessing it.

Let

p denote the proportion of sampling units in the population belonging to class 1, and

q the proportion of units falling in class 2.

Evidently, Np will be the number of sampling units in the population belonging to class 1, Nq the number of sampling units in class 2, and $Np + Nq = N$. Now, clearly, the probability $P(n_1)$ that in a sample of n selected out of N by the method of simple random sampling, n_1 will occur in class 1 and n_2 in class 2 will be given by

$$P(n_1) = \binom{n}{n_1} \left\{\frac{Np}{N} \cdot \frac{Np-1}{N-1} \cdots \frac{Np-n_1+1}{N-n_1+1}\right\}$$

$$\times \left\{\frac{Nq}{N-n_1} \cdot \frac{Nq-1}{N-n_1-1} \cdots \frac{Nq-n_2+1}{N-n+1}\right\}$$

which can also be alternatively written as

$$P(n_1) = \frac{\binom{Np}{n_1}\binom{Nq}{n_2}}{\binom{N}{n}} \tag{68}$$

The variate n_1 or the proportion n_1/n is said to be distributed in a *hyper-geometric* distribution. Since the possible values which n_1 can assume are 0, 1, ..., n, we have

$$\sum_{n_1=0}^{n} P(n_1) = 1$$

or

$$\sum_{n_1=0}^{n} \frac{\binom{Np}{n_1}\binom{Nq}{n_2}}{\binom{N}{n}} = 1 \tag{69}$$

As N tends to be large, the distribution (68) approaches the binomial, the probability of observing n_1 in class 1 and n_2 in class 2 in a sample of n being now given by

$$\binom{n}{n_1} p^{n_1}(1-p)^{n_2}$$

2*a*.15 Mean Value of the Hyper-Geometric Distribution

By definition,

$$E(n_1) = \sum_{n_1=0}^{n} n_1 P(n_1)$$

Substituting from (68) for $P(n_1)$, we have

$$E(n_1) = \sum_{n_1=0}^{n} n_1 \frac{Np!}{n_1!\,(Np-n_1)!}\;\frac{Nq!}{n_2!\,(Nq-n_2)!}\;\frac{n!\,(N-n)!}{N!}$$

$$= \frac{Nnp}{N}\sum_{n_1=1}^{n} \frac{(Np-1)!}{(n_1-1)!\,(Np-n_1)!}\cdot\frac{Nq!}{n_2!\,(Nq-n_2)!}$$

$$\times\ \frac{(n-1)!\,(N-n)!}{(N-1)!}$$

$$= np \sum_{n_1=1}^{n} \frac{\binom{Np-1}{n_1-1}\binom{Nq}{n_2}}{\binom{N-1}{n-1}}$$

Now

$$\frac{\binom{Np-1}{n_1-1}\binom{Nq}{n_2}}{\binom{N-1}{n-1}}$$

represents the probability that in a sample of $n-1$, n_1-1 will fall in class 1 and n_2 will fall in class 2. Consequently

$$\sum_{n_1=1}^{n} \frac{\binom{Np-1}{n_1-1}\binom{Nq}{n_2}}{\binom{N-1}{n-1}} = 1$$

Hence

$$E(n_1) = np \tag{70}$$

Or, denoting by p_n the proportion in the sample, we write

$$E(p_n) = p \tag{71}$$

It follows that an unbiased estimate of the proportion p in the population is given by the proportion in the sample. In other words,

$$\text{Est. } p = \frac{n_1}{n} = p_n \tag{72}$$

Similarly,

$$\text{Est. } q = \frac{n_2}{n} = q_n \tag{73}$$

2*a*.16 Variance of the Hyper-Geometric Distribution

By definition,

$$V(n_1) = E(n_1^2) - \{E(n_1)\}^2$$

Now

$$n_1^2 = n_1(n_1 - 1) + n_1 \tag{74}$$

so that

$$V(n_1) = E\{n_1(n_1-1)\} + E(n_1) - \{E(n_1)\}^2 \tag{75}$$

Also

$$\begin{aligned}E\{n_1(n_1-1)\} &= \sum_{n_1=1}^{n} n_1(n_1-1)P(n_1)\\ &= \sum_{n_1=1}^{n} n_1(n_1-1)\frac{Np!}{n_1!(Np-n_1)!}\,\frac{Nq!}{n_2!(Nq-n_2)!}\\ &\qquad\qquad\times\frac{n!(N-n)!}{N!}\\ &= \frac{n(n-1)Np(Np-1)}{N(N-1)}\sum_{n_1=2}^{n}\frac{(Np-2)!}{(n_1-2)!(Np-n_1)!}\\ &\qquad\qquad\times\frac{Nq!}{n_2!(Nq-n_2)!}\,\frac{(n-2)!(N-n)!}{(N-2)!}\\ &= \frac{n(n-1)Np(Np-1)}{N(N-1)} \end{aligned}\tag{76}$$

since the sum of the terms under the summation sign is evidently 1. Substituting the result in (75), we have

$$\begin{aligned}V(n_1) &= \frac{n(n-1)Np(Np-1)}{N(N-1)} + np - n^2p^2\\ &= \frac{N-n}{N-1}\,npq\end{aligned}\tag{77}$$

It follows that the sampling variance of the estimated proportion is given by

$$V(p_n) = \frac{N-n}{N-1}\cdot\frac{pq}{n} \tag{78}$$

and the standard error of the estimated proportion is given by

$$S.E.(p_n) = \sqrt{\frac{N-n}{N-1}\cdot\frac{pq}{n}} \tag{79}$$

These results can also be obtained directly from those of the preceding sections. All that one need do is to adopt the convention of scoring the character of a sampling unit with one whenever it appears in class 1 and zero when it falls in class 2. On making these substitutions, we obtain

$$\bar{y}_N = \frac{1}{N}\sum_{i=1}^{N} y_i = \frac{Np}{N} = p \tag{80}$$

$$\bar{y}_n = \frac{1}{n}\sum^{n} y_i = \frac{n_1}{n} = p_n \tag{81}$$

$$S^2 = \frac{\sum\limits_{i=1}^{N} y_i^2 - N\bar{y}_N^2}{N-1} = \frac{Np - Np^2}{N-1} = \frac{N}{N-1}\, p\,(1-p) \tag{82}$$

$$s^2 = \frac{\sum\limits^{n} y_i^2 - n\bar{y}_n^2}{n-1} = \frac{n_1 - \frac{n_1^2}{n}}{n-1} = \frac{n}{n-1}\, p_n\,(1-p_n) \tag{83}$$

On substituting the above in (16), we reach the result (71); and on substitution in (39), we get the expression for the variance of the sample proportion, as in (78).

Further, from (36), we get

$$E\left\{\frac{n}{n-1}\, p_n\,(1-p_n)\right\} = \frac{N}{N-1}\, p\,(1-p) \tag{84}$$

It follows from (84) that an unbiased estimate of the product $p\,(1-p)$ is given by

$$\frac{n}{n-1}\;\frac{N-1}{N}\; p_n\,(1-p_n)$$

and not by

$$p_n\,(1-p_n)$$

as one might suppose. We write

$$\text{Est. }\{p\,(1-p)\} = \frac{n\,(N-1)}{(n-1)\,N}\cdot p_n\,(1-p_n) \tag{85}$$

The unbiased estimate of the sampling variance of a proportion in terms of the sample values is, therefore, given by

$$\text{Est. } V(p_n) = \frac{N-n}{N}\;\frac{p_n\,(1-p_n)}{n-1}. \tag{86}$$

and an estimate of the standard error of p_n is given by

$$\text{Est.}\,S.E.\,(p_n) = \sqrt{\frac{N-n}{N}\,\frac{p_n\,(1-p_n)}{n-1}} \tag{87}$$

This estimate, however, has a negative bias.

2a.17 Confidence Limits and Size of Sample for Specified Precision

The confidence limits for the proportion are derived on the same assumptions as for the quantitative characters, namely, that the sample proportion p is normally distributed. This will approximately be so, unless p is too small (or large) and n is small. The limits are given by

$$p = p_n \pm t_{(\alpha,\,\infty)} \sqrt{\frac{N-n}{N-1}\,\frac{p\,(1-p)}{n}} \tag{88}$$

where $t_{(\alpha,\,\infty)}$ denotes, as before, the value of t corresponding to the significance level α and ∞ degrees of freedom. This can be solved as a quadratic in p (Bartlett, 1937) or, alternatively, as a quicker approximation by substituting from (85) for $p\,(1-p)$, giving

$$p = p_n \pm t_{(\alpha,\,\infty)} \sqrt{\frac{N-n}{N}\,\frac{p_n\,(1-p_n)}{n-1}} \tag{89}$$

As p deviates from $\cdot 5$, the distribution of p_n becomes remote from the normal, and the normal theory ceases to be applicable unless n and N are both large. The appropriate method in this case is to obtain the confidence limits directly from the hypergeometric distribution itself. The probability of getting n_1 or fewer numbers in class 1 in a sample of n, is the sum of the corresponding terms of the hyper-geometric series. Equating this sum to the chosen level of significance, it is possible to solve for p. For any larger values of p, the probability would be smaller than α, and for smaller values greater than α. The value of p given by this equation will, therefore, give an upper limit. Similarly, by equating α to the sum of the terms of the hyper-geometric series for variates greater than or equal to n_2, we obtain the lower limit to p. The method, however, involves heavy computations.

The size of sample required for estimating the population proportion with a specified precision is obtained from (88). If the error permissible in the estimate is, say, ϵp and the degree of assurance desired is $1 - \alpha$, we then have

$$n = \frac{\dfrac{t^2_{(\alpha,\,\infty)}\,q}{\epsilon^2 p}}{1 + \dfrac{1}{N}\left\{\dfrac{t^2_{(\alpha,\,\infty)}q}{\epsilon^2 p} - 1\right\}} \tag{90}$$

For N large and ϵ not too small, n is simply given by

$$n = \frac{t^2_{(\alpha,\,\infty)}\,q}{\epsilon^2 p} \tag{91}$$

Example 2.1

Material for the construction of 5000 wells was issued during the year 1944 in a certain district as part of the Grow-More-Food Campaign in India. The list of cultivators to whom it was issued, together with the proposed location of each well, is available. A large part of the material was reported to have been misused by diverting it to other purposes. It is proposed to assess the extent of the misuse by means of a sample spot check. In other words, it is proposed to estimate the proportion of wells actually constructed and used for irrigation purposes. The sample is proposed to be selected by the method of simple random sampling from the total population of wells for which the material was issued. The permissible margin of error in the estimated value is 10% and the degree of assurance desired is 95%. Determine the size of sample for values of p ranging from ·5 to ·9.

We are given $N = 5000$, $\epsilon = \cdot 10$ and $t_{(\alpha,\,\infty)} = 1\cdot 960$.

Substituting in (90), we obtain for different values of p, the following values of n.

p	..	·5	·6	·7	·8	·9
n	..	357	244	159	94	42

Since the worst critics do not place the misuse at more than half of the material issued, a sample of 357 would appear to be adequate for the proposed check.

2*a*.18 Generalised Hyper-Geometric Distribution

We shall now extend the preceding results to the population which is divided into k mutually exclusive classes. Let N_i denote the number of units in the i-th class of the population ($i = 1, 2, \ldots, k$), so that

$$\sum_{i=1}^{k} N_i = N$$

Now, if a simple random sample of n is drawn out of this population then it can be seen by analogy with the distribution of two classes that the probability that n_i units occur in the i-th class and $n - n_i$ in all the other $k - 1$ classes together is given by

$$P\{n_i\} = \frac{\binom{N_i}{n_i}\binom{N - N_i}{n - n_i}}{\binom{N}{n}}$$

More generally, it can be seen that the probability that n_1 units occur in class 1, n_2 in class 2, ..., and n_k in class k, is given by

$$P\{n_1, n_2, \ldots, n_k\} = \frac{\binom{N_1}{n_1}\binom{N_2}{n_2}\cdots\binom{N_k}{n_k}}{\binom{N}{n}} \tag{92}$$

It follows from (70) and (77) that

$$E(n_i) = np_i \tag{93}$$

where

$$p_i = \frac{N_i}{N}$$

and

$$V(n_i) = \frac{N-n}{N-1} \cdot np_i(1 - p_i) \tag{94}$$

It should be pointed out that the numbers falling in any two classes are not independent of each other, since

$$\sum_{i=1}^{k} n_i = n$$

$$P\{n_{11}, n_{12}, n - n_{11} - n_{12}\} = \frac{\binom{N_{11}}{n_{11}} \binom{N_{12}}{n_{12}} \binom{N - N_{11} - N_{12}}{n - n_{11} - n_{12}}}{\binom{N}{n}} \tag{102}$$

The probability that in a sample of n, $n_{11} + n_{12}$ (*i.e.*, n_1) will fall in classes 1 and 2 together and $n - n_{11} - n_{12}$ (*i.e.*, $n - n_1$) in classes 3 and 4 taken together is similarly given by

$$P\{n_{11} + n_{12}, n - n_{11} - n_{12}\} = \frac{\binom{N_{11} + N_{12}}{n_{11} + n_{12}} \binom{N - N_{11} - N_{12}}{n - n_{11} - n_{12}}}{\binom{N}{n}} \tag{103}$$

We know by a well-known theorem in probabiiity that

$$P\{n_{11}, n_{12}, n - n_1\} = P\{n_1, n - n_1\} \; P\{n_{11}, n_{12} \mid n_1\} \tag{104}$$

It follows that the probability required is given by the quotient of (102) by (103) and is equal to

$$P\{n_{11}, n_{12} | \, n_1\} = \frac{\binom{N_{11}}{n_{11}} \binom{N_{12}}{n_{12}}}{\binom{N_1}{n_1}} \tag{105}$$

In other words, n_{11}/n_1 follows a hyper-geometric distribution in samples of n_1 drawn from N_1. Hence, from (101), we obtain

$$E\left\{\frac{n_{11}}{n_1} \middle| \, n_1\right\} = \frac{N_{11}}{N_1} = \frac{p_{11}}{p_1} \tag{106}$$

where $p_{11} = N_{11}/N$, $p_1 = N_1/N$. The result is otherwise obvious also, for, with n_1 fixed, the situation reduces to the two-class problem and we apply the results of Section 2*a*.15. Substituting the result in (100), we have

$$\begin{aligned} E\left(\frac{n_{11}}{n_1}\right) &= E\left(\frac{N_{11}}{N_1}\right) \\ &= \frac{N_{11}}{N_1} \\ &= \frac{p_{11}}{p_1} \end{aligned} \tag{107}$$

To obtain the variance of n_{11}/n_1, we proceed similarly. By definition,

$$V\left(\frac{n_{11}}{n_1}\right) = E\left\{\left(\frac{n_{11}}{n_1}\right)^2\right\} - \left\{E\left(\frac{n_{11}}{n_1}\right)\right\}^2$$

$$= E\left[E\left\{\frac{n_{11}^2}{n_1^2}\,\middle|\, n_1\right\}\right] - \frac{N_{11}^2}{N_1^2}$$

$$= E\left[E\left\{\frac{n_{11}(n_{11}-1)+n_{11}}{n_1^2}\,\middle|\, n_1\right\}\right] - \frac{N_{11}^2}{N_1^2} \qquad (108)$$

We have seen that n_{11}/n_1 follows a hyper-geometric distribution in samples of n_1 drawn from N_1. It follows that

$$E\left\{\frac{n_{11}}{n_1^2}\,\middle|\, n_1\right\} = \frac{N_{11}}{N_1}\,\frac{1}{n_1} = \frac{p_{11}}{p_1}\,\frac{1}{n_1} \qquad (109)$$

and from (76) we have

$$E\left\{\frac{n_{11}(n_{11}-1)}{n_1^2}\,\middle|\, n_1\right\} = \frac{(n_1-1)\,N_{11}\,(N_{11}-1)}{n_1\,N_1\,(N_1-1)} \qquad (110)$$

Substituting in (108), we have

$$V\left(\frac{n_{11}}{n_1}\right) = E\left[\left(1-\frac{1}{n_1}\right)\frac{N_{11}\,(N_{11}-1)}{N_1\,(N_1-1)} + \frac{1}{n_1}\,\frac{N_{11}}{N_1}\right] - \frac{N_{11}^2}{N_1^2}$$

$$= \frac{N_{11}\,(N_{11}-1)}{N_1\,(N_1-1)} + \frac{N_{11}}{N_1}\left(1-\frac{N_{11}-1}{N_1-1}\right)E\left(\frac{1}{n_1}\right) - \frac{N_{11}^2}{N_1^2}$$

$$= -\frac{N_{11}}{N_1}\left\{\frac{N_1-N_{11}}{N_1\,(N_1-1)}\right\} + \frac{N_{11}}{N_1}\left(\frac{N_1-N_{11}}{N_1-1}\right)E\left(\frac{1}{n_1}\right)$$

$$= \frac{N_{11}\,(N_1-N_{11})}{N_1\,(N_1-1)}\left\{E\left(\frac{1}{n_1}\right) - \frac{1}{N_1}\right\} \qquad (111)$$

Now let

$$n_1 = np_1 + \epsilon$$

where

$$E(\epsilon) = 0$$

and

$$E(\epsilon^2) = \frac{N-n}{N-1}\cdot n\cdot p_1\,(1-p_1)$$

Since ϵ will be small as compared with np_1, with a probability approaching 1 as n becomes large, we may write, neglecting terms in $1/n^2$ and higher powers,

$$\frac{1}{n_1} \cong \frac{1}{np_1}\left\{1 - \frac{\epsilon}{np_1} + \frac{\epsilon^2}{n^2 p_1{}^2} - \cdots\right\}$$

Hence

$$E\left(\frac{1}{n_1}\right) \cong \frac{1}{np_1}$$

to a first approximation, or more precisely

$$E\left(\frac{1}{n_1}\right) \cong \frac{1}{np_1}\left\{1 + \frac{N-n}{N-1}\,\frac{1}{n}\,\frac{1}{p_1}\,(1-p_1)\right\} \tag{112}$$

Substituting from (112) in (111), we have

$$V\left(\frac{n_{11}}{n_1}\right) = \frac{p_{11}}{p_1}\cdot\frac{N_1 - N_{11}}{N_1 - 1}\left\{\frac{1}{np_1} - \frac{1}{Np_1} + \frac{N-n}{N-1}\,\frac{1}{n^2 p_1{}^2}(1-p_1)\right\} \tag{113}$$

Replacing $N_1 - 1$, $N - 1$ by N_1, N respectively, we get

$$V_1\left(\frac{n_{11}}{n_1}\right) \cong \frac{N-n}{N}\cdot\frac{1}{n}\cdot\frac{1}{p_1}\cdot\frac{p_{11}}{p_1}\left(1 - \frac{p_{11}}{p_1}\right) \tag{114}$$

where V_1 denotes the first approximation to the variance, and

$$V_2\left(\frac{n_{11}}{n_1}\right) \cong \frac{N-n}{N}\cdot\frac{1}{np_1}\cdot\frac{p_{11}}{p_1}\left(1 - \frac{p_{11}}{p_1}\right)\left\{1 + \frac{1}{np_1}(1-p_1)\right\} \tag{115}$$

where V_2 denotes the second approximation to the variance.

2a.20 Quantitative and Qualitative Characters

We will now extend the preceding theory to the situation involving both quantitative and qualitative variation together in the same problem. This situation is of common occurrence. Thus, in a population survey we may be required to estimate both the proportions of families in different income groups as also the total income in each group. The tabulation on a sampling basis of census results presents similar problems. Suppose punched cards, each one representing the data of different holdings in an agricultural census, are available for sorting and tabulation. Further, suppose that the holdings are to be classified according to their size in five classes: 0–2·5, 2·5–5·0, 5–10, 10–25 and larger than 25 acres. We may be required to estimate proportions of holdings in the different classes and also the total area under any specified crop in each class. In all such problems, it is convenient to select a sample of n out of the total of N by the method of simple random sampling. We have already considered the problem of estimating the proportions in the different classes. The problems for consideration now are:

(*a*) to obtain an estimate of the total (or the average) of the quantitative character under study in each class;

(*b*) to obtain the standard error of the estimates in (*a*); and

(*c*) to predict the size of the sample n required for estimating the total in each class with a given standard error.

Without loss of generality, we may consider these problems in relation to an actual example.

Example 2.2

It is proposed to estimate the area benefited from irrigation wells said to have been completed under the Grow-More-Food Campaign from the data given in Example 2.1. The sample is proposed to be selected by the method of simple random sampling from the population of wells reported to have been constructed. The number of wells actually constructed is not known. How large should n be in order that the area benefited may be determined with 5% standard error?

Let

N = Number of wells reported to have been constructed under the Grow-More-Food Campaign;

p = Proportion of wells in the population actually constructed; for convenience we will designate these wells as belonging to class 1; and

q = Proportion of wells not completed. We will designate the wells under this category as falling in class 2.

Evidently,

$$Np + Nq = N$$

Let, further, $\bar{y}_{N_1}$, S_1^2 be respectively the population mean area benefited per well and the population mean square for class 1.

Let n_1 denote the number of wells falling in class 1 when a random sample of n is chosen by the method of simple random sampling from the population N, and $\bar{y}_{n_1}$ be the corresponding mean area in the sample. Our first problem is to obtain the estimate of the total area benefited, namely, $Np\bar{y}_{N_1}$.

Since the sample is chosen by the method of simple random sampling from the entire population, the sub-sample n_1 can also be considered a random sample from the corresponding population of Np units. It follows that for a given n_1, $\bar{y}_{n_1}$ will be an unbiased estimate of $\bar{y}_{N_1}$.

It is, therefore, natural to take $N . n_1/n . \bar{y}_{n_1}$ as the estimate of the total area benefited from the completed wells. It is easy to show that this is an unbiased estimate of $Np\bar{y}_{N_1}$. For

$$\begin{aligned} E\left(N \cdot \frac{n_1}{n} \cdot \bar{y}_{n_1}\right) &= E\left\{E\left(\frac{N}{n} \cdot n_1 \cdot \bar{y}_{n_1} \mid n_1\right)\right\} \\ &= E\left(\frac{N}{n}\, n_1\, \bar{y}_{N_1}\right) \\ &= \frac{N}{n}\, \bar{y}_{N_1}\, E(n_1) \\ &= \frac{N}{n}\, \bar{y}_{N_1}\, np \\ &= Np\bar{y}_{N_1} \end{aligned} \tag{116}$$

The next problem is to obtain the sampling variance of $N \cdot n_1/n \cdot \bar{y}_{n_1}$. We have

$$\begin{aligned} V\left\{\frac{N}{n} \cdot n_1\bar{y}_{n_1}\right\} &= \frac{N^2}{n^2}\, V\left\{\sum^{n_1} y_i\right\} \\ &= \frac{N^2}{n^2}\left\{E\left[\left(\sum^{n_1} y_i\right)^2\right] - \left[E\left(\sum^{n_1} y_i\right)\right]^2\right\} \\ &= \frac{N^2}{n^2}\, E\left[\left(\sum^{n_1} y_i\right)^2\right] - N^2p^2\bar{y}_{N_1}^2 \end{aligned} \tag{117}$$

Now, for given n_1,

$$E\left[\left\{\sum^{n_1} y_i\right\}^2 \middle| \, n_1\right] = E\left\{\sum^{n_1} y_i^2 \middle| \, n_1\right\} + E\left\{\sum_{i \neq j}^{n_1} y_i y_j \middle| \, n_1\right\}$$

$$= \frac{n_1}{Np} \sum_{i=1}^{Np} y_i^2 + \frac{n_1(n_1-1)}{Np(Np-1)} \sum_{i \neq j=1}^{Np} y_i y_j \tag{118}$$

Hence

$$E\left[\left\{\sum^{n_1} y_i\right\}^2\right] = E\left[E\left\{\left(\sum^{n_1} y_i\right)^2 \middle| \, n_1\right\}\right]$$

$$= \frac{1}{Np}\left(\sum_{i=1}^{Np} y_i^2\right) E(n_1)$$

$$+ \frac{1}{Np(Np-1)} \left(\sum_{i \neq j=1}^{Np} y_i y_j\right) E(n_1 \cdot \overline{n_1 - 1}) \tag{119}$$

Using the results, already established, namely,

$$E(n_1) = np$$

and

$$E\ (n_1 \cdot \overline{n_1 - 1}) = \frac{Np(Np-1)}{N(N-1)}\, n(n-1)$$

we obtain from (119)

$$E\left[\left\{\sum^{n_1} y_i\right\}^2\right] = \frac{n}{N} \sum_{i=1}^{Np} y_i^2 + \frac{n(n-1)}{N(N-1)} \sum_{i \neq j=1}^{Np} y_i y_j$$

$$= \frac{n}{N} \sum_{i=1}^{Np} y_i^2 + \frac{n(n-1)}{N(N-1)} \left\{\left(\sum_{i=1}^{Np} y_i\right)^2 - \sum_{i=1}^{Np} y_i^2\right\}$$

$$= \frac{n}{N}\ \frac{N-n}{N-1} \left(\sum_{i=1}^{Np} y_i^2\right) + \frac{n(n-1)}{N(N-1)} N^2 p^2 \bar{y}_{N_1}{}^2$$

$$= \frac{n}{N}\ \frac{N-n}{N-1} \{(Np-1)\,S_1^2 + Np\bar{y}_{N_1}{}^2\}$$

$$+ \frac{n(n-1)}{N(N-1)} N^2 p^2 \bar{y}_{N_1}{}^2 \tag{120}$$

Substituting the result in (117), we have

$$V\left\{\frac{N}{n}\cdot n_1\bar{y}_{n_1}\right\} = \frac{N(N-n)}{n(N-1)}\{(Np-1)\,S_1^2 + Np\bar{y}_{N_1}{}^2\}$$

$$+\frac{N(n-1)}{n(N-1)}N^2p^2\bar{y}_{N_1}{}^2 - N^2p^2\bar{y}_{N_1}{}^2$$

$$= \frac{N(N-n)}{n(N-1)}(Np-1)\,S_1^2$$

$$+ N^2p\bar{y}_{N_1}{}^2\left\{\frac{N-n}{n(N-1)} + \frac{Np(n-1)}{n(N-1)} - p\right\}$$

$$= \frac{N(N-n)}{n(N-1)}(Np-1)\,S_1^2 + \frac{N^2(N-n)\,p\,(1-p)}{n(N-1)}\bar{y}_{N_1}{}^2 \qquad (121)$$

An alternative and instructive way of deriving the above result is to proceed as follows. We have

$$n_1\bar{y}_{n_1} = \Sigma^{n_1} y_i$$

$$= \Sigma^{n} y_i$$

since $n - n_1$ of the y values are zero each,

$$= n\bar{y}_n$$

It follows, therefore, that

$$V\left\{\frac{N}{n}\cdot n_1\bar{y}_{n_1}\right\} = V\{N\bar{y}_n\}$$

$$= N^2 . \frac{N-n}{N} . \frac{S^2}{n}$$

$$= \frac{N(N-n)}{n(N-1)}\left\{\sum_{i=1}^{N} y_i^2 - N\bar{y}_N{}^2\right\}$$

$$= \frac{N(N-n)}{n(N-1)}\left\{\sum_{i=1}^{Np} y_i^2 - \frac{(Np)^2}{N}\bar{y}_{N_1}{}^2\right\}$$

$$= \frac{N(N-n)}{n(N-1)}\{(Np-1)\,S_1^2 + Np(1-p)\,\bar{y}_{N_1}{}^2\}$$

For purposes of simplification, we will assume that Np is large enough to permit the following approximations:

$$\frac{Np-1}{Np} \cong 1$$

and

$$\frac{N-1}{N} \cong 1$$

Using these, we obtain

$$V\left\{\frac{N}{n} \cdot n_1 \bar{y}_{n_1}\right\} = \frac{N(N-n)}{n}\{p\mathrm{S}_1^2 + p(1-p)\,\bar{y}_{N_1}^2\} \tag{122}$$

To predict the size of sample required for estimating the character with a given standard error, we need the expression for the relative variance. This is obtained by dividing (122) by $N^2p^2\bar{y}_{N_1}^2$ and is given by

$$\frac{N-n}{N} \cdot \frac{1}{n}\left\{\frac{C_1^2}{p} + \frac{1-p}{p}\right\} \tag{123}$$

where C_1^2 denotes the square of the coefficient of variation of the area irrigated from a well in class 1. For N large, the relative variance is given by

$$\frac{1}{n}\left\{\frac{C_1^2}{p} + \frac{1-p}{p}\right\} \tag{124}$$

An idea of C_1^2 may be formed from previous experience. Let us assume it to be 0·5. Since n will need to be large as p decreases, p may be assumed to be the smallest of the values consistent with expectation and previous experience in order that we may err on the safe side. Table 2.3 gives values of n for different values of p and for $C_1^2 = 0{\cdot}5$ in order that the area benefited may be estimated with 5% standard error.

Two sets of values of n are given:

(i) those obtained from (123), for $N = 5000$, and

(ii) those from (124), *i.e.*, after neglecting the finite multiplier.

It will be seen that a sample of 690 wells will be required for estimating the area benefited with a degree of accuracy as large as the one specified or larger, assuming of course that p does not fall below $\cdot 5$. Ignoring the finite multiplier altogether would imply a loss of nearly 15% of the information.

TABLE 2.3

Sample Size Required for Estimating the Total in Class 1 with 5% Standard Error

p		·5	·6	·7	·8	·9
(i)	..	690	536	419	327	253
(ii)	..	800	600	457	350	267

We may call attention to one important point. It will be seen from (122) that sampling units falling in a given class alone contribute to the information in that class. It follows that this formula for the sampling variance is applicable to any class, even when the population consists of several classes, p in that case representing the proportion of units in the population falling in the given class, and $(1 - p)$ representing the proportion of units falling in all the remaining classes together. It also follows that the value of n required for estimating the class areas with a specified accuracy or higher is the value corresponding to the smallest of the p values.

B. SAMPLING WITH VARYING PROBABILITIES OF SELECTION

2*b*.1 Introduction

The theory considered in the previous sections is appropriate for the method of simple random sampling in which the selection probabilities are equal for all units in the population. Although this method is by far the most common method of sampling, unequal probabilities of selection are sometimes used, and give more efficient estimates in the sense of giving population estimates with smaller standard errors. In the following sections we shall give the basic theory of this method.

There is one difference between the method of simple random sampling and that of sampling with varying probabilities of selection. In the former, the probability of drawing a specified unit remains constant at each draw; in the latter, it does not. Let, as before, P_i denote the probability of selecting the i-th unit of the population at the first draw ($i = 1, 2, \ldots, N$), so that $\sum_{i=1}^{N} P_i = 1$ and P_{i_r} denote the probability of drawing y_i at the r-th draw ($r = 1, 2, \ldots, n$).

Clearly,

$$P_{i_1} = P_i \qquad (i = 1, 2, \ldots, N) \tag{125}$$

and

$$\begin{aligned}
P_{i_2} &= \text{(Probability that } y_i \text{ is not drawn at the first draw)} \times \\
&\quad \text{(Probability that } y_i \text{ is drawn at the second draw)} \\
&= \sum_{j(\neq i)=1}^{N} \text{(Probability that } y_j \text{ is drawn at the first draw)} \times \\
&\quad \text{(Probability that } y_i \text{ is drawn at the second draw)} \\
&= \left\{P_1 \cdot \frac{P_i}{1-P_1} + P_2 \cdot \frac{P_i}{1-P_2} + \cdots + P_{i-1} \cdot \frac{P_i}{1-P_{i-1}} \right. \\
&\quad \left. + P_{i+1} \cdot \frac{P_i}{1-P_{i+1}} + \cdots + P_N \cdot \frac{P_i}{1-P_N}\right\} \\
&= \left\{\sum_{i=1}^{N} \frac{P_i}{1-P_i} - \frac{P_i}{1-P_i}\right\} P_i \\
&= \left\{S - \frac{P_i}{1-P_i}\right\} P_i
\end{aligned} \tag{126}$$

where

$$S = \sum_{i=1}^{N} \frac{P_i}{1-P_i} \tag{127}$$

It is thus seen that P_{i_2} is not equal to P_{i_1} unless $P_i = 1/N$. The theory of sampling with varying probabilities of selection is consequently more complex than that of simple random sampling. One way of introducing simplification into the theory is to replace a selected unit before another draw is made, so that $P_{i_r} = P_i$

for all r. We shall first present this simplified theory appropriate to the procedure of sampling with replacement and then consider briefly the theory appropriate for sampling without replacement.

2*b*.2 Sampling with Replacement: Sample Estimate and its Variance

Define a variate z given by

$$z_i = \frac{y_i}{NP_i} \tag{128}$$

and consider the simple arithmetic mean of z values in the sample given by

$$\bar{z}_n = \frac{1}{n} \sum^{n} z_i \tag{129}$$

It is easily shown that $\bar{z}_n$ provides an unbiased estimate of the population mean $\bar{y}_N$.

For, in sampling with replacement,

$$E(z_i) = \sum_{i=1}^{N} P_i z_i \tag{130}$$

which we may denote by $\bar{z}_{..}$, at each draw.

Substituting for z_i from (128) on the right-hand side of (130), we get

$$\begin{aligned} E(z_i) &= \sum_{i=1}^{N} P_i \cdot \frac{y_i}{NP_i} \\ &= \frac{1}{N} \sum_{i=1}^{N} y_i \\ &= \bar{y}_N \end{aligned} \tag{131}$$

It follows that

$$\begin{aligned} E(\bar{z}_n) &= \frac{1}{n} \sum^{n} E(z_i) \\ &= \bar{y}_N \end{aligned} \tag{132}$$

To obtain the sampling variance of $\bar{z}_n$, we have

$$\begin{aligned} V(\bar{z}_n) &= E\{\bar{z}_n - E(\bar{z}_n)\}^2 \\ &= E(\bar{z}_n^2) - \{E(\bar{z}_n)\}^2 \\ &= E\left\{\frac{1}{n}\sum^{n} z_i\right\}^2 - \bar{z}_{..}^2 \\ &= \frac{1}{n^2}\, E\left\{\sum^{n} z_i^2 + \sum_{i\neq j}^{n} z_i z_j\right\} - \bar{z}_{..}^2 \\ &= \frac{1}{n^2}\left\{\sum^{n} E(z_i^2) + \sum_{i\neq j}^{n} E(z_i z_j)\right\} - \bar{z}_{..}^2 \qquad (133) \end{aligned}$$

Now

$$E(z_i^2) = \sum_{i=1}^{N} P_i z_i^2 \qquad (134)$$

and

$$E(z_i z_j) = E(z_i) \cdot E(z_j)$$

since draws are made with replacement.

On substituting for $E(z_i)$ and $E(z_j)$ from (130), we have

$$\begin{aligned} E(z_i z_j) &= \left(\sum_{i=1}^{N} P_i z_i\right)\left(\sum_{j=1}^{N} P_j z_j\right) \\ &= \bar{z}_{..}^2 \qquad (135) \end{aligned}$$

Substituting from (134) and (135) in (133), we obtain

$$\begin{aligned} V(\bar{z}_n) &= \frac{1}{n^2}\left\{n \sum_{i=1}^{N} P_i z_i^2 + n(n-1)\,\bar{z}_{..}^2\right\} - \bar{z}_{..}^2 \\ &= \frac{1}{n}\left\{\sum_{i=1}^{N} P_i z_i^2 - \bar{z}_{..}^2\right\} \\ &= \frac{1}{n}\left\{\sum_{i=1}^{N} P_i (z_i - \bar{z}_{..})^2\right\} \\ &= \frac{\sigma_z^2}{n} \qquad (136) \end{aligned}$$

where

$$\sigma_z^2 = \sum_{i=1}^{N} P_i \, (z_i - \bar{z}_{..})^2 \tag{137}$$

and represents the variance of a single z.

It will be noticed that the finite multiplier does not enter into the expression for the variance of the estimate when sampling is carried out with replacement. When $P_i = 1/N$, we have

$$\bar{z}_n = \bar{y}_n$$

and

$$\sigma_z^2 = \frac{1}{N} \sum_{i=1}^{N} (y_i - \bar{y}_N)^2$$

$$= \sigma_y^2$$

and $V(\bar{y}_n)$ is given by the familiar expression

$$V(\bar{y}_n) = \frac{\sigma_y^2}{n}$$

$$= \frac{N-1}{N} \cdot \frac{S^2}{n} \tag{138}$$

Lastly, we remark that when the selection probability is proportional to the value of the variate, in other words, when P_i is proportional to y_i, say, $P_i = y_i/\mu$, z assumes a constant value for all i, and in consequence (136) reduces to zero. In practice ıhe values of the variate will, of course, not be known in advance but values of another variate correlated with the variate under study may be known. We may, therefore, expect that when P_i is proportional to the measure of size of y_i, the estimate may be considerably more efficient than that based on simple random sampling.

2*b*.3 Sampling with Replacement: Estimation of the Sampling Variance

Consider the mean square of z_i's obtained from the sample, defined by

$$s_z^2 = \frac{1}{n-1} \sum^{n} (z_i - \bar{z}_n)^2 \tag{139}$$

Expanding, and taking expectations, we get

$$E(s_z^2) = \frac{1}{n-1} E\left\{\sum^n z_i^2 - n\bar{z}_n^2\right\}$$

$$= \frac{1}{n-1}\left\{\sum^n E(z_i^2) - nE(\bar{z}_n^2)\right\} \tag{140}$$

Now, by definition,

$$V(\bar{z}_n) = E(\bar{z}_n^2) - \bar{z}_{..}^2$$

so that

$$E(\bar{z}_n^2) = V(\bar{z}_n) + \bar{z}_{..}^2$$

$$= \frac{\sigma_z^2}{n} + \bar{z}_{..}^2 \tag{141}$$

Substituting from (134) and (141) in (140), we get

$$E(s_z^2) = \frac{1}{n-1}\left\{n\sum_{i=1}^{N} P_i z_i^2 - \sigma_z^2 - n\bar{z}_{..}^2\right\}$$

$$= \sigma_z^2 \tag{142}$$

thus showing that s_z^2 is an unbiased estimate of σ_z^2. It follows that

$$\text{Est. } V(\bar{z}_n) = \frac{s_z^2}{n} \tag{143}$$

2b.4 Sampling without Replacement

We have already noted that in sampling without replacement, when selection probabilities are unequal, the probability of drawing a specified unit of the population at a given draw is not in general equal to the probability of selecting it at the first draw. It follows that the expected value of a variate will change with the successive draws. In this section we shall consider the theory of sampling without replacement for the simple case when the sample consists of only two draws.

One device of overcoming the difficulty of changing expectation with the successive draws is to define z differently for the successive draws. Thus, let

$$z_1' = \frac{y_i}{NP_{i_1}} \tag{144}$$

and

$$z_2' = \frac{y_j}{NP_{j_2}}$$

$$= \frac{y_j}{N\left(S - \frac{P_j}{1-P_j}\right) P_j} \tag{145}$$

where y_i and y_j are values of the units drawn at the first and second draws respectively.

Clearly,

$$E(z_1') = \sum_{i=1}^{N} P_{i_1} \cdot \frac{y_i}{NP_{i_1}}$$

$$= \bar{y}_N \tag{146}$$

Also,

$$E(z_2') = \sum_{j=1}^{N} P_{j_2} \cdot \frac{y_j}{N\left(S - \frac{P_j}{1-P_j}\right) P_j}$$

$$= \bar{y}_N \tag{147}$$

It follows that a simple arithmetic mean of z_1' and z_2' will provide us with an unbiased estimate of the population mean.

An alternative estimate, in which z is defined independently of the order of the draw and which is unbiased, can also be formed. Let

$$z_i = \frac{2}{N} \cdot \frac{y_i}{\left(S + 1 - \frac{P_i}{1-P_i}\right) P_i} \tag{148}$$

and consider the estimate

$$\bar{z}_{(n=2)} = \tfrac{1}{2} \sum^{2} z_i \tag{149}$$

Clearly,

$$E(\bar{z}_{(n=2)}) = \tfrac{1}{2} \sum_{i=1}^{N} E(a_i)\, z_i \tag{150}$$

where $E(a_i)$ stands for the probability of including y_i in a sample of two, and is given by

$$\begin{aligned} E(a_i) &= \text{Probability that } y_i \text{ is drawn at the first draw} \\ &\quad + \text{Probability that } y_i \text{ is drawn at the second draw} \end{aligned}$$

Using (125) and (126), we have

$$\begin{aligned} E(a_i) &= P_i + \left\{S - \frac{P_i}{1-P_i}\right\} P_i \\ &= \left\{S + 1 - \frac{P_i}{1-P_i}\right\} P_i \end{aligned} \tag{151}$$

Substituting for $E(a_i)$ from (151) in (150), we get

$$\begin{aligned} E(\bar{z}_{(n=2)}) &= \tfrac{1}{2} \sum_{i=1}^{N} \left\{S + 1 - \frac{P_i}{1-P_i}\right\} P_i z_i \\ &= \bar{z}_{..}', \text{ say} \end{aligned} \tag{152}$$

Substituting for z_i from (148) in (152), we have

$$\begin{aligned} E(\bar{z}_{(n=2)}) &= \tfrac{1}{2} \sum_{i=1}^{N} \frac{2}{N} \cdot \frac{y_i}{\left(S + 1 - \frac{P_i}{1-P_i}\right) P_i} \\ &\qquad\qquad \times \left(S + 1 - \frac{P_i}{1-P_i}\right) P_i \\ &= \bar{y}_N \end{aligned} \tag{153}$$

thus showing that $\bar{z}_{(n=2)}$ provides an unbiased estimate of the population mean $\bar{y}_N$.

We shall next obtain the variance of $\bar{z}_{(n=2)}$. We write

$$\begin{aligned} V(\bar{z}_{(n=2)}) &= E(\bar{z}^2_{(n=2)}) - \bar{y}_N^2 \\ &= \frac{1}{n^2} E\left\{\sum^{n} z_i^2 + \sum_{i \neq j}^{n} z_i z_j\right\} - \bar{y}_N^2 \\ &= \frac{1}{n^2} \left\{\sum_{i=1}^{N} E(a_i)\, z_i^2 + \sum_{i \neq j=1}^{N} E(a_i a_j)\, z_i z_j\right\} - \bar{y}_N^2 \end{aligned} \tag{154}$$

where $E(a_i a_j)$ stands for the probability of including y_i and y_j in a sample of two and is given by

$$E(a_i a_j) = P_{i_1} P_{j_2 | i} + P_{j_1} P_{i_2 | j}$$
$$= P_i \cdot \frac{P_j}{1 - P_i} + P_j \cdot \frac{P_i}{1 - P_j} \tag{155}$$

On substituting for $E(a_i)$ from (151) and for $E(a_i a_j)$ from (155) in (154), we have

$$V(\bar{z}_{(n=2)}) = \tfrac{1}{4}\left[\sum_{i=1}^{N}\left\{S + 1 - \frac{P_i}{1 - P_i}\right\} P_i z_i^2 + \sum_{i \neq j = 1}^{N} P_i P_j \left\{\frac{1}{1 - P_i} + \frac{1}{1 - P_j}\right\} z_i z_j\right] - \bar{y}_N^2 \tag{156}$$

To obtain the estimate of the variance, we have from the definition of the variance

$$\text{Est. } V(\bar{z}_{(n=2)}) = \bar{z}^2_{(n=2)} - \text{Est. } \bar{z}_{..}'^2 \tag{157}$$

Now, from (152),

$$\bar{z}_{..}'^2 = \tfrac{1}{4}\left\{\sum_{i=1}^{N}\left(S + 1 - \frac{P_i}{1 - P_i}\right)^2 P_i^2 z_i^2 + \sum_{i \neq j = 1}^{N}\left(S + 1 - \frac{P_i}{1 - P_i}\right)\left(S + 1 - \frac{P_j}{1 - P_j}\right) P_i P_j z_i z_j\right\} \tag{158}$$

It is easily shown that

$$E\left\{\sum^{n}\left(S + 1 - \frac{P_i}{1 - P_i}\right) P_i z_i^2\right\}$$
$$= \sum_{i=1}^{N} E(a_i)\left(S + 1 - \frac{P_i}{1 - P_i}\right) P_i z_i^2$$
$$= \sum_{i=1}^{N}\left(S + 1 - \frac{P_i}{1 - P_i}\right)^2 P_i^2 z_i^2 \tag{159}$$

Also,

$$E\left\{\sum_{i\neq j}^{n}\left(S+1-\frac{P_i}{1-P_i}\right)\left(S+1-\frac{P_j}{1-P_j}\right) \times \frac{z_i z_j}{\dfrac{1}{1-P_i}+\dfrac{1}{1-P_j}}\right\}$$

$$=\sum_{i\neq j=1}^{N} E\,(a_i a_j)\left(S+1-\frac{P_i}{1-P_i}\right)\left(S+1-\frac{P_j}{1-P_j}\right) \times \frac{z_i z_j}{\dfrac{1}{1-P_i}+\dfrac{1}{1-P_j}}$$

$$=\sum_{i\neq j=1}^{N}\left(S+1-\frac{P_i}{1-P_i}\right)\left(S+1-\frac{P_j}{1-P_j}\right) P_i P_j\, z_i z_j \tag{160}$$

It follows from (157), (158), (159) and (160) that

$$\text{Est. } V\,(\bar{z}_{(n=2)}) = \bar{z}^2_{(n=2)} - \tfrac{1}{4}\left\{\sum^{n}\left(S+1-\frac{P_i}{1-P_i}\right) P_i z_i^2 + \sum_{i\neq j}^{n}\frac{\left(S+1-\dfrac{P_i}{1-P_i}\right)\left(S+1-\dfrac{P_j}{1-P_j}\right)}{\dfrac{1}{1-P_i}+\dfrac{1}{1-P_j}}\, z_i z_j\right\}$$

$$= \bar{z}^2_{(n=2)} - \frac{1}{N^2}\sum^{n}\frac{y_i^2}{\left(S+1-\dfrac{P_i}{1-P_i}\right)P_i} - \frac{1}{N^2}\sum_{i\neq j}^{n}\frac{y_i y_j}{P_i P_j\left(\dfrac{1}{1-P_i}+\dfrac{1}{1-P_j}\right)} \tag{161}$$

2b.5 Sampling without Replacement—General Case

The formal general expression for an unbiased estimate of $\bar{y}_N$, for any sample size, should now be obvious.

Let

$$z_i = \frac{ny_i}{NE\,(a_i)} \tag{162}$$

Then the simple arithmetic mean of z's provides an unbiased estimate of $\bar{y}_N$, for,

$$\begin{aligned} E\,(\bar{z}_n) &= E\left\{\frac{1}{n}\sum^{n}\frac{ny_i}{NE\,(a_i)}\right\} \\ &= \frac{1}{n}\sum_{i=1}^{N} E\,(a_i)\cdot\frac{ny_i}{NE\,(a_i)} \\ &= \bar{y}_N \end{aligned} \tag{163}$$

To obtain the sampling variance, we write

$$\begin{aligned} V\,(\bar{z}_n) &= E\,(\bar{z}_n{}^2) - \bar{y}_N{}^2 \\ &= \frac{1}{n^2}\,E\left\{\sum^{n} z_i{}^2 + \sum_{i\neq j}^{n} z_i z_j\right\} - \bar{y}_N{}^2 \\ &= \frac{1}{n^2}\left\{\sum_{i=1}^{N} E\,(a_i)\,z_i{}^2 + \sum_{i\neq j=1}^{N} E\,(a_i a_j)\,z_i z_j\right\} - \bar{y}_N{}^2 \end{aligned} \tag{164}$$

The expression appears to have been first given by Narain (1951) and Horvitz and Thompson (1952).

Substituting for z_i from (162) in (164), we have

$$V\,(\bar{z}_n) = \frac{1}{N^2}\left\{\sum_{i=1}^{N}\frac{y_i{}^2}{E\,(a_i)} + \sum_{i\neq j=1}^{N}\frac{E\,(a_i a_j)}{E\,(a_i)\,E\,(a_j)}\,y_i y_j\right\} - \bar{y}_N{}^2 \tag{165}$$

which for $n = 2$ reduces to (156).

An estimate of the variance $V\,(\bar{z}_n)$ is easily derived. We write

$$\text{Est. } V\,(\bar{z}_n) = \bar{z}_n{}^2 - \text{Est. } \bar{y}_N{}^2$$

$$= \bar{z}_n^2 - \frac{1}{N^2} \text{ Est.} \left\{ \sum_{i=1}^{N} y_i^2 + \sum_{i \neq j = 1}^{N} y_i y_j \right\}$$

$$= \bar{z}_n^2 - \frac{1}{N^2} \sum^{n} \frac{y_i^2}{E(\alpha_i)} - \frac{1}{N^2} \sum_{i \neq j}^{n} \frac{y_i y_j}{E(\alpha_i \alpha_j)} \qquad (166)^*$$

The values of $E(\alpha_i)$ and $E(\alpha_i\alpha_j)$ depend upon the choice of values for $P_{i_r} \binom{i = 1, 2, \ldots, N}{r = 1, 2, \ldots, n}$. Explicit expressions have been given for $n = 2$ in the last section and expressions could be written in a similar way for any sample size. It will be noticed that if $E(\alpha_i)$ is proportional to y_i, the estimate $\bar{z}_n$ reduces to a constant with zero variance. Values of y_i will, however, not be known but values A_i of a character correlated with y_i may be known. It follows as a near approximation that the optimum values of P_{i_r} would require that $E(\alpha_i)$ should be proportional to A_i. The explicit solution of this problem is, however, difficult and will not be discussed here. One result may be mentioned, namely, that it is possible to choose $E(\alpha_i)$ proportional to A_i only if the latter are not too heterogeneous (Narain, 1951).

* Yates and Grundy (1953) have developed an alternative estimate of the variance $V(\bar{z}_n)$, which appears to be better than the one given in (166). They use the fact that

$$\sum_{j(\neq i)=1}^{N} \{E(\alpha_i\alpha_j) - E(\alpha_i)E(\alpha_j)\} = -E(\alpha_i)\{1 - E(\alpha_i)\}$$

to recast the expression for the variance $V(\bar{z}_n)$ as a linear function of the squares of the differences of the z's. Thus, (165) can be written as

$$V(\bar{z}_n) = \frac{1}{N^2} \left[\sum_{i=1}^{N} \frac{1 - E(\alpha_i)}{E(\alpha_i)} y_i^2 + \sum_{i \neq j=1}^{N} \frac{E(\alpha_i\alpha_j) - E(\alpha_i)E(\alpha_j)}{E(\alpha_i)E(\alpha_j)} y_i y_j \right]$$

Substituting for y_i in terms of z_i from (162), and using the above result, they obtain

$$V(\bar{z}_n) = \frac{1}{2n^2} \sum_{i \neq j=1}^{N} \{E(\alpha_i)E(\alpha_j) - E(\alpha_i\alpha_j)\}(z_i - z_j)^2$$

It is easy to see that an unbiased estimate of $V(\bar{z}_n)$ is given by

$$\text{Est. } V(\bar{z}_n) = \frac{1}{2n^2} \sum_{i \neq j}^{n} \left\{ \frac{E(\alpha_i)E(\alpha_j) - E(\alpha_i\alpha_j)}{E(\alpha_i\alpha_j)} \right\} (z_i - z_j)^2$$

which, unlike (166), is seen to be a linear function of the squares of the differences between the z's in the sample.

Lastly, we mention one system of selection probabilities due to Midzuno (1950) which, while by no means efficient in the sense indicated above, has the merit of simplicity. Here the units at the first draw are selected with unequal probability, but at the second and subsequent draws they are selected with equal probability.

It is easy to see that under this system

$$\begin{aligned} E(a_i) &= P_{i_1} + P_{i_2} + P_{i_3} + \cdots + P_{i_n} \\ &= P_{i_1} + (1 - P_{i_1}) \frac{1}{N-1} + (1 - P_{i_1} - P_{i_2}) \frac{1}{N-2} + \cdots \\ &= P_i + (1 - P_i) \frac{1}{N-1} + \left\{1 - P_i - (1 - P_i) \frac{1}{N-1}\right\} \\ &\qquad \times \frac{1}{N-2} + \cdots \\ &= P_i + \frac{n-1}{N-1} (1 - P_i) \\ &= \frac{N-n}{N-1} P_i + \frac{n-1}{N-1} \end{aligned} \tag{167}$$

while

$$\begin{aligned} E\left(a_i a_j \,\middle|\, \begin{matrix} n \\ N \end{matrix}\right) &= P_{i_1} E(a_j \mid a_i) + P_{j_1} E(a_i \mid a_j) \\ &\qquad + (1 - P_{i_1} - P_{j_1}) \cdot E\left(a_i a_j \,\middle|\, \begin{matrix} n-1 \\ N-1 \end{matrix}\right) \\ &= P_i \frac{n-1}{N-1} + P_j \frac{n-1}{N-1} + (1 - P_i - P_j) \\ &\qquad \times \frac{(n-1)(n-2)}{(N-1)(N-2)} \\ &= \frac{n-1}{N-1} \left\{\frac{N-n}{N-2} (P_i + P_j) + \frac{n-2}{N-2}\right\} \end{aligned} \tag{168}$$

Incidentally, we note that under this system the probability of drawing a specified sample is proportional to the total measure of the units included in the sample.

REFERENCES

1. David, F. N. and Neyman, J. (1938) .. "Extension of the Markoff Theorem on Least Squares," *Statist. Res. Mem.*, **2**, 105–16.

2. Sukhatme, P. V. (1938) .. "On Bipartitional Functions," *Phil. Trans. Roy. Soc., London*, Series A, **237**, 375–409.

3. ——— (1944) .. "Moments and Product Moments of Moment-Statistics for Samples of the Finite and Infinite Populations," *Sankhya*, **6**, 363–82.

4. Fisher, R. A. and Yates, F. (1938) *Statistical Tables for Biological, Agricultural and Medical Research*, Oliver and Boyd, Ltd., London.

5. Neyman, J. (1934) .. "On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection," *Jour. Roy. Statist. Soc.*, **97**, 558–606.

6. Sukhatme, P. V. (1935) .. "Contribution to the Theory of the Representative Method," *Jour. Roy. Statist. Soc. Suppl.*, **2**, 253–68.

7. Bartlett, M. S. (1937) .. "Sub-sampling for Attributes," *Jour. Roy. Statist. Soc. Suppl.*, **4**, 131–35.

8. Narain, R. D. (1951) .. "On Sampling without Replacement with Varying Probabilities," *Jour. Ind. Soc. Agr. Statist.*, **3**, 169–74.

9. Horvitz, D. G. and Thompson, D. J. (1952) "A Generalisation of Sampling without Replacement from a Finite Universe," *Jour. Amer. Statist. Assoc.*, **47**, 663–85.

10. Midzuno, H. (1950) .. "An Outline of the Theory of Sampling Systems," *Ann. Inst. Statist. Math., Japan*, **1**, 149–56.

APPENDIX

Tables of $s_g\,(P, Q)/\chi_1!\,\chi_2!\ldots$

$w = 1$

		Q
		(1)
P	(1)	1

$w = 2$

		Q	
		(2)	(1^2)
P	(2)	1	.
	(1^2)	1	1

$w = 3$

		Q		
		(3)	(21)	(1^3)
	(3)	1	.	.
P	(21)	1	1	.
	(1^3)	1	3	1

$w = 4$

				Q		
		(4)	(31)	(2^2)	(21^2)	(1^4)
	(4)	1	.	.	.	.
	(31)	1	1	.	.	.
P	(2^2)	1	.	1	.	.
	(21^2)	1	2	1	1	.
	(1^4)	1	4	3	6	1

$w = 5$

					Q			
		(5)	(41)	(32)	(31^2)	(2^21)	(21^3)	(1^5)
	(5)	1	.	.	.	.	.	.
	(41)	1	1	.	.	.	.	.
	(32)	1	.	1	.	.	.	.
P	(31^2)	1	2	1	1	.	.	.
	(2^21)	1	1	2	.	1	.	.
	(21^3)	1	3	4	3	3	1	.
	(1^5)	1	5	10	10	15	10	1

Tables of $s_g\,(P, Q)/\chi_1!\,\chi_2!\ldots$

$w = 6$

	Q										
	(6)	(51)	(42)	(3^2)	(41^2)	(321)	(2^3)	(31^3)	(2^21^2)	(21^4)	(1^6)
(6)	1	.	.	.	.	.	.	.	.	.	.
(51)	1	1	.	.	.	.	.	.	.	.	.
(42)	1	.	1	.	.	.	.	.	.	.	.
(3^2)	1	.	.	1	.	.	.	.	.	.	.
(41^2)	1	2	1	.	1	.	.	.	.	.	.
P (321)	1	1	1	1	.	1	.	.	.	.	.
(2^3)	1	.	3	.	.	.	1	.	.	.	.
(31^3)	1	3	3	1	3	3	.	1	.	.	.
(2^21^2)	1	2	3	2	1	4	1	.	1	.	.
(21^4)	1	4	7	4	6	16	3	4	6	1	.
(1^6)	1	6	15	10	15	60	15	20	45	15	1

Tables of $s_g\ (P, Q)/\chi_1!\ \chi_2! \ldots$

$w = 8$

	Q																					
P	(8)	(71)	(62)	(53)	(4^2)	(61^2)	(521)	(42^2)	(431)	(3^22)	(51^3)	(421^2)	(3^21^2)	(32^21)	(2^4)	(41^4)	(321^3)	(2^31^2)	(31^5)	(2^21^4)	(21^6)	(1^8)
(8)	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
(71)	1	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
(62)	1	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
(53)	1	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
(4^2)	1	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
(61^2)	1	2	1	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
(521)	1	1	1	1	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
(42^2)	1	.	2	.	1	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.
(431)	1	1	.	1	1	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.
(3^22)	1	.	1	2	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.
(51^3)	1	3	3	1	.	3	3	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.
(421^2)	1	2	2	2	1	1	2	1	2	.	.	1	.	.	.	.	.	.	.	.	.	.
(3^21^2)	1	2	1	2	2	1	.	.	4	1	.	.	1	.	.	.	.	.	.	.	.	.
(32^21)	1	1	2	3	1	.	2	1	1	2	.	.	.	1	.	.	.	.	.	.	.	.
(2^4)	1	.	4	.	3	.	.	6	.	.	.	.	.	.	1	.	.	.	.	.	.	.
(41^4)	1	4	6	4	1	6	12	3	4	.	4	6	.	.	.	1	.	.	.	.	.	.
(321^3)	1	3	4	5	3	3	6	3	9	4	1	3	3	3	.	.	1	.	.	.	.	.
(2^31^2)	1	2	4	6	3	1	6	6	6	6	.	3	.	6	1	.	.	1	.	.	.	.
(31^5)	1	5	10	11	5	10	30	15	25	10	10	30	10	15	.	5	10	.	1	.	.	.
(2^21^4)	1	4	8	12	7	6	20	16	28	20	4	18	12	28	3	1	8	6	.	1	.	.
(21^6)	1	6	16	26	15	15	66	60	90	70	20	105	60	150	15	15	80	45	6	15	1	.
(1^8)	1	8	28	56	35	28	168	210	280	280	56	420	280	840	105	70	560	420	56	210	28	1

Tables of s_g $(P, Q)/\chi_1!\,\chi_2!\ldots$

$w = 7$

								Q							
		(7)	(61)	(52)	(43)	(51^2)	(421)	(3^21)	(32^2)	(41^3)	(321^2)	(2^31)	(31^4)	(2^21^3)	(21^5)	(1^7)
	(7)	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.
	(61)	1	1	.	.	.	.	.	.	.	.	.	.	.	.	.
	(52)	1	.	1	.	.	.	.	.	.	.	.	.	.	.	.
	(43)	1	.	.	1	.	.	.	.	.	.	.	.	.	.	.
	(51^2)	1	2	1	.	1	.	.	.	.	.	.	.	.	.	.
	(421)	1	1	1	1	.	1	.	.	.	.	.	.	.	.	.
	(3^21)	1	1	.	2	.	.	1	.	.	.	.	.	.	.	.
P	(32^2)	1	.	2	1	.	.	.	1	.	.	.	.	.	.	.
	(41^3)	1	3	3	1	3	3	.	.	1	.	.	.	.	.	.
	(321^2)	1	2	2	3	1	2	2	1	.	1	.	.	.	.	.
	(2^31)	1	1	3	3	.	3	.	3	.	.	1	.	.	.	.
	(31^4)	1	4	6	5	6	12	4	3	4	6	.	1	.	.	.
	(2^21^3)	1	3	5	7	3	9	6	7	1	6	3	.	1	.	.
	(21^5)	1	5	11	15	10	35	20	25	10	40	15	5	10	1	.
	(1^7)	1	7	21	35	21	105	70	105	35	210	105	35	105	21	1

Tables of $\pi_1!\,\pi_2!\ldots g_s\,(P, Q)$

$w = 1$			$w = 2$				$w = 3$				
		Q			Q					Q	
		(1)			(2)	(1^2)			(3)	(21)	(1^3)
P	(1)	1	P	(2)	1	.		(3)	1	.	.
							P	(21)	−1	1	.
				(1^2)	−1	1		(1^3)	2	−3	1

$w = 4$

				Q		
		(4)	(31)	(2^2)	(21^2)	(1^4)
	(4)	1	.	.	.	.
	(31)	−1	1	.	.	.
P	(2^2)	−1	.	1	.	.
	(21^2)	2	−2	−1	1	.
	(1^4)	−6	8	3	−6	1

$w = 5$

					Q			
		(5)	(41)	(32)	(31^2)	(2^21)	(21^3)	(1^5)
	(5)	1	.	.	.	.	.	.
	(41)	−1	1	.	.	.	.	.
	(32)	−1	.	1	.	.	.	.
P	(31^2)	2	− 2	− 1	1	.	.	.
	(2^21)	2	− 1	− 2	.	1	.	.
	(21^3)	−6	6	5	−3	−3	1	.
	(1^5)	24	−30	−20	20	15	−10	1

Tables of $\pi_1!\,\pi_2!\ldots g_s\,(P, Q)$

$w = 6$

		Q										
		(6)	(51)	(42)	(3^2)	(41^2)	(321)	(2^3)	(31^3)	(2^21^2)	(21^4)	(1^6)
	(6)	1	.	.	.	.	.	.	.	.	.	.
	(51)	− 1	1	.	.	.	.	.	.	.	.	.
	(42)	− 1	.	1	.	.	.	.	.	.	.	.
	(3^2)	− 1	.	.	1	.	.	.	.	.	.	.
	(41^2)	2	− 2	− 1	.	1	.	.	.	.	.	.
P	(321)	2	− 1	− 1	− 1	.	1	.	.	.	.	.
	(2^3)	2	.	− 3	.	.	.	1	.	.	.	.
	(31^3)	− 6	6	3	2	− 3	− 3	.	1	.	.	.
	(2^21^2)	− 6	4	5	2	− 1	− 4	− 1	.	1	.	.
	(21^4)	24	−24	−18	− 8	12	20	3	−4	− 6	1	.
	(1^6)	−120	144	90	40	−90	−120	−15	40	45	−15	1

Tables of $\pi_1!\,\pi_2!\ldots g_S\,(P, Q)$

$w = 7$

								Q							
	(7)	(61)	(52)	(43)	(51^2)	(421)	(3^21)	(32^2)	(41^3)	(321^2)	(2^31)	(31^4)	(2^21^3)	(21^5)	(1^7)
(7)	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.
(61)	−1	1	.	.	.	.	.	.	.	.	.	.	.	.	.
(52)	−1	.	1	.	.	.	.	.	.	.	.	.	.	.	.
(43)	−1	.	.	1	.	.	.	.	.	.	.	.	.	.	.
(51^2)	2	−2	−1	.	1	.	.	.	.	.	.	.	.	.	.
(421)	2	−1	−1	−1	.	1	.	.	.	.	.	.	.	.	.
(3^21)	2	−1	.	−2	.	.	1	.	.	.	.	.	.	.	.
P (32^2)	2	.	−2	−1	.	.	.	1	.	.	.	.	.	.	.
(41^3)	−6	6	3	2	−3	−3	.	.	1	.	.	.	.	.	.
(321^2)	−6	4	3	4	−1	−2	−2	−1	.	1	.	.	.	.	.
(2^31)	−6	2	6	3	.	−3	.	−3	.	.	1	.	.	.	.
(31^4)	24	−24	−12	−14	12	12	8	3	−4	−6	.	1	.	.	.
(2^21^3)	24	−18	−18	−14	6	15	6	8	−1	−6	−3	.	1	.	.
(21^5)	−120	120	84	70	−60	−90	−40	−35	20	50	15	−5	−10	1	.
(1^7)	720	−840	−504	−420	504	630	280	210	−210	−420	−105	70	105	−21	1

Tables of $\pi_1!\,\pi_2!\ldots g_s\,(P, Q)$

$w = 8$

							Q					
		(8)	(71)	(62)	(53)	(4^2)	(61^2)	(521)	(42^2)	(431)	(3^22)	(51^3)
	(8)	1	.	.	.	.	.	.	.	.	.	.
	(71)	−1	1	.	.	.	.	.	.	.	.	.
	(62)	−1	.	1	.	.	.	.	.	.	.	.
	(53)	−1	.	.	1	.	.	.	.	.	.	.
	(4^2)	−1	.	.	.	1	.	.	.	.	.	.
	(61^2)	2	−2	−1	.	.	1	.	.	.	.	.
	(521)	2	−1	−1	−1	.	.	1	.	.	.	.
	(42^2)	2	.	−2	.	−1	.	.	1	.	.	.
	(431)	2	−1	.	−1	−1	.	.	.	1	.	.
	(3^22)	2	.	−1	−2	.	.	.	.	.	1	.
P	(51^3)	−6	6	3	2	.	−3	−3	.	.	.	1
	(421^2)	−6	4	3	2	2	−1	−2	−1	−2	.	.
	(3^21^2)	−6	4	1	4	2	−1	.	.	−4	−1	.
	(32^21)	−6	2	4	4	1	.	−2	−1	−1	−2	.
	(2^4)	−6	.	8	.	3	.	.	−6	.	.	.
	(41^4)	24	−24	−12	−8	−6	12	12	3	8	.	−4
	(321^3)	24	−18	−12	−14	−6	6	9	3	12	5	−1
	(2^31^2)	24	−12	−20	−12	−6	2	12	9	6	6	.
	(31^5)	−120	120	60	64	30	−60	−60	−15	−70	−20	20
	(2^21^4)	−120	96	84	64	30	−36	−72	−33	−56	−28	8
	(21^6)	720	−720	−480	−384	−180	360	504	180	420	160	−120
	(1^8)	−5040	5760	3360	2688	1260	−3360	−4032	−1260	−3360	−1120	1344

Tables of $\pi_1!\,\pi_2!\,\ldots g_s\,(P, Q)$—*(Contd.)*

$w = 8$

						Q					
	(421^2)	(3^21^2)	(32^21)	(2^4)	(41^4)	(321^3)	(2^31^2)	(31^5)	(2^21^4)	(21^6)	(1^8)
(8)	.	.	.	.	.	.	.	.	.	.	.
(71)	.	.	.	.	.	.	.	.	.	.	.
(62)	.	.	.	.	.	.	.	.	.	.	.
(53)	.	.	.	.	.	.	.	.	.	.	.
(4^2)	.	.	.	.	.	.	.	.	.	.	.
(61^2)	.	.	.	.	.	.	.	.	.	.	.
(521)	.	.	.	.	.	.	.	.	.	.	.
(42^2)	.	.	.	.	.	.	.	.	.	.	.
(431)	.	.	.	.	.	.	.	.	.	.	.
(3^22)	.	.	.	.	.	.	.	.	.	.	.
P (51^3)	.	.	.	.	.	.	.	.	.	.	.
(421^2)	1	.	.	.	.	.	.	.	.	.	.
(3^21^2)	.	1	.	.	.	.	.	.	.	.	.
(32^21)	.	.	1	.	.	.	.	.	.	.	.
(2^4)	.	.	.	1	.	.	.	.	.	.	.
(41^4)	− 6	.	.	.	1	.	.	.	.	.	.
(321^3)	− 3	− 3	− 3	.	.	1	.	.	.	.	.
(2^31^2)	− 3	.	− 6	− 1	.	.	1	.	.	.	.
(31^5)	30	20	15	.	− 5	− 10	.	1	.	.	.
(2^21^4)	30	12	32	3	− 1	− 8	− 6	.	1	.	.
(21^6)	−270	−120	−210	−15	30	100	45	−6	−15	1	.
(1^8)	2520	1120	1680	105	−420	−1120	−420	112	210	−28	1

CHAPTER III

STRATIFIED SAMPLING

A. SELECTION WITH EQUAL PROBABILITY

3*a*.1 Introduction

We have seen that the precision of a sample estimate of the population mean depends upon two factors: (1) the size of the sample, and (2) the variability or heterogeneity of the population. Apart from the size of the sample, therefore, the only way of increasing the precision of an estimate is to devise sampling procedures which will effectively reduce the heterogeneity. One such procedure is known as the procedure of stratified sampling. It consists in dividing the population into k classes and drawing random samples of known sizes, one each from the different classes. The classes into which the population is divided are called the *strata* and the process is termed the procedure of *stratified sampling* as distinct from the procedure considered in the previous chapters, called unrestricted or *unstratified sampling*. An example of stratified sampling is furnished by the survey for estimating the average yield of a crop per acre in which administrative areas are taken as the strata and random samples of predetermined numbers of fields are selected from each of the several strata. The geographical proximity of fields within a stratum makes it more homogeneous than the entire population and thus helps to increase the precision of the estimate. In this chapter we shall consider the theory applicable to the procedure of stratified sampling.

Stratified sampling is a common procedure in sample surveys. The procedure ensures any desired representation in the sample of all the strata in the population. In unstratified sampling, on the other hand, adequate representation of all the strata cannot always be ensured and indeed a sample may be so distributed among the different strata that certain strata may be over-represented and others under-represented. The procedure of stratified sampling is thus intended to give a better cross-section of the population than that of unstratified sampling. It follows that one would expect the precision of the estimated character to be higher in stratified than in unstratified sampling. Stratified sampling also serves other purposes. The selection of sampling units, the location and enumeration of the selected units and the distribution and supervision of field work are all simplified in stratified sampling. Of course, stratified sampling presupposes the knowledge of the strata sizes, *i.e.*, the total number of sampling units in each stratum and the availability of the frame for the selection of the sample from each stratum.

It is not necessary that the strata be formed of geographically contiguous administrative areas. Thus, in yield surveys, the fields may be stratified according as they are irrigated or unirrigated and separate samples selected from each. In a survey for estimating the acreage under crops, strata may be formed by classifying the villages according to their geographical area instead of on the basis of geographical proximity. The principles to be followed in stratifying a population will become clear in the subsequent sections.

3*a*.2 Estimate of the Population Mean and its Variance

Let N_i denote the size, *i.e.*, the number of units in the *i*-th stratum and n_i the size of the sample to be selected therefrom, so that

$$\sum_{i=1}^{k} N_i = N$$

and

$$\sum_{i=1}^{k} n_i = n$$

Now the population mean to be estimated is given by

$$\begin{aligned}\bar{y}_N &= \frac{1}{N}\sum_{i=1}^{k} N_i\bar{y}_{N_i} \\ &= \sum_{i=1}^{k} p_i\bar{y}_{N_i} \qquad (1)\end{aligned}$$

Since n_i is a simple random sample from N_i, it is natural to take

$$\sum_{i=1}^{k} p_i\bar{y}_{n_i} \qquad (2)$$

as an estimate of the population mean and denote it by $\bar{y}_w$ as it is the weighted mean with strata sizes as the weights. It is easy to see that this gives an unbiased etsimate of the population mean, since

$$E(\bar{y}_{n_i}) = \bar{y}_{N_i}$$

and, therefore,

$$E(\bar{y}_w) = E\left\{\sum_{i=1}^{k} p_i\bar{y}_{n_i}\right\} = \sum_{i=1}^{k} p_i E(\bar{y}_{n_i}) = \sum_{i=1}^{k} p_i\bar{y}_{N_i} = \bar{y}_N \tag{3}$$

To obtain the sampling variance of $\bar{y}_w$, we have

$$\begin{aligned} V(\bar{y}_w) &= E(\bar{y}_w - \bar{y}_N)^2 \\ &= E\left(\sum_{i=1}^{k} p_i\bar{y}_{n_i} - \sum_{i=1}^{k} p_i\bar{y}_{N_i}\right)^2 \\ &= E\left\{\sum_{i=1}^{k} p_i(\bar{y}_{n_i} - \bar{y}_{N_i})\right\}^2 \\ &= E\left\{\sum_{i=1}^{k} p_i^2(\bar{y}_{n_i} - \bar{y}_{N_i})^2 + \sum_{i\neq j=1}^{k} p_i p_j(\bar{y}_{n_i} - \bar{y}_{N_i})(\bar{y}_{n_j} - \bar{y}_{N_j})\right\} \\ &= \sum_{i=1}^{k} p_i^2 E(\bar{y}_{n_i} - \bar{y}_{N_i})^2 + \sum_{i\neq j=1}^{k} p_i p_j E\{(\bar{y}_{n_i} - \bar{y}_{N_i})(\bar{y}_{n_j} - \bar{y}_{N_j})\} \quad (4) \end{aligned}$$

Since the sample in the i-th stratum is a simple random sample,

$$E(\bar{y}_{n_i} - \bar{y}_{N_i})^2 = \left(\frac{1}{n_i} - \frac{1}{N_i}\right)\mathbf{S}_i^2 \tag{5}$$

where $\mathbf{S}_i^2$ is the mean square of the population in the i-th stratum defined by

$$\mathbf{S}_i^2 = \frac{1}{N_i - 1}\sum_{j=1}^{N_i}(y_{ij} - \bar{y}_{N_i})^2 \tag{6}$$

The value of the second term in (4) is clearly zero, since samples are selected independently from each stratum. We therefore have

$$V(\bar{y}_w)_S = \sum_{i=1}^{k}\left(\frac{1}{n_i} - \frac{1}{N_i}\right)p_i^2\mathbf{S}_i^2 \tag{7}$$

The subscript '*S*' in (7) helps to indicate that this variance relates to stratified sampling and will be used whenever necessary. It will be seen that the variance depends on S_i's, the variabilities within the strata. The result suggests that the smaller the S_i, *i.e.*, the more homogeneous the strata are made, the greater will be the precision of the stratified sample.

It can be shown by a slight modification of the Markoff theorem mentioned in Section 2*a*.3 that $\bar{y}_w$ is the best unbiased linear estimate of $\bar{y}_N$. The proof is left to the reader.

3*a*.3 Choice of Sample Sizes in Different Strata

The above expression for the variance of the estimated mean in stratified sampling shows that the precision of a stratified sample depends upon n_i's which can be fixed at will. Following the principle explained in Section 1.3, we can now so choose the n_i's as to provide an estimate of the desired precision for a minimum cost. Alternatively, for any given cost, we can choose the n_i's so as to minimize the variance of the estimate. The allocation of the sample to the different strata made in accordance with either of these principles is said to be based on the principle of *optimum allocation*. The concept, as it is known to-day, was introduced by Neyman (1934).

Just as the variance is a function of the sample sizes, so also is the cost of a survey. The manner in which the cost will vary with the size of the total sample and with its allocation among the different strata will depend upon the particular survey. In the simplest case, such as in sampling from punched cards for tabulating the results of a census on a sampling basis, the cost will be directly proportional to the number of units in the sample. In yield surveys in India, where the field work is carried out by the local staff in the course of their normal duties and the major item in the cost of a survey consists of labour charges on harvesting of produce, the cost of the survey is found to be approximately proportional to the number of crop-cutting experiments. Cost per experiment may, however, vary in the different strata depending upon the availability of labour. In this situation, the total cost may be appropriately represented by

$$C = \sum_{i=1}^{k} c_i n_i \tag{8}$$

where c_i is the cost per experiment in the i-th stratum. When c_i is the same from stratum to stratum, say c, the total cost of a survey is given by

$$C = cn \tag{9}$$

Where travel, field staff salary and statistical analysis are to be paid for, the cost function will obviously change in form. We shall later on give examples of more complex cost functions appropriate to different problems.

To determine the optimum values of n_i, when the cost function is represented by (8), we consider the function

$$\phi = V(\bar{y}_w) + \mu C$$

where μ is some constant, and note that

$$\begin{aligned} V(\bar{y}_w) + \mu C &= \sum_{i=1}^{k} \left(\frac{1}{n_i} - \frac{1}{N_i}\right) p_i^2 \mathbf{S}_i^2 + \mu \left\{\sum_{i=1}^{k} c_i n_i\right\} \\ &= \sum_{i=1}^{k} \left\{\left(\frac{p_i^2 \mathbf{S}_i^2}{n_i} + \mu c_i n_i - 2 p_i \mathbf{S}_i \sqrt{\mu c_i}\right) \right. \\ &\qquad\qquad \left. + 2 p_i \mathbf{S}_i \sqrt{\mu c_i} - \frac{1}{N_i} p_i^2 \mathbf{S}_i^2 \right\} \\ &= \sum_{i=1}^{k} \left(\frac{p_i \mathbf{S}_i}{\sqrt{n_i}} - \sqrt{\mu c_i n_i}\right)^2 + 2 \sum_{i=1}^{k} p_i \mathbf{S}_i \sqrt{\mu c_i} \\ &\qquad\qquad - \sum_{i=1}^{k} \frac{1}{N_i} p_i^2 \mathbf{S}_i^2 \\ &= \sum_{i=1}^{k} \left(\frac{p_i \mathbf{S}_i}{\sqrt{n_i}} - \sqrt{\mu c_i n_i}\right)^2 \\ &\qquad\qquad + \text{terms independent of } n_i \end{aligned} \tag{10}$$

Clearly, $V(\bar{y}_w)$ is minimum for fixed C, or the cost C of a survey is minimum for a fixed value of $V(\bar{y}_w)$, when each of the square terms on the right-hand side of (10) is zero; or in other words, when

$$n_i \propto \frac{p_i \mathbf{S}_i}{\sqrt{c_i}} \qquad (i = 1, 2, \ldots, k) \tag{11}$$

Equation (11) shows, what is also otherwise obvious, that:

(i) the larger the size of the stratum, the larger should be the size of the sample to be selected therefrom;

(ii) the larger the variability within a stratum, the larger should be the size of the sample from that stratum; and

(iii) the cheaper the labour in a stratum, the larger the sample from that stratum.

To obtain the exact value of n_i's, we evaluate $1/\sqrt{\mu}$, the constant of proportionality, so as to satisfy the condition of fixed cost or fixed variance. In the former case, we substitute from (11) in (8) and obtain

$$C_0 = \sum_{i=1}^{k} \frac{p_i S_i}{\sqrt{\mu c_i}} c_i$$

where C_0 is the budgeted amount within which it is desired to estimate the mean with the maximum precision, giving

$$\frac{1}{\sqrt{\mu}} = \frac{C_0}{\sum\limits_{i=1}^{k} p_i S_i \sqrt{c_i}} \tag{12}$$

Hence

$$n_i = \frac{p_i S_i C_0}{\sqrt{c_i}\left(\sum\limits_{i=1}^{k} p_i S_i \sqrt{c_i}\right)} \tag{13}$$

When $c_i = c$ $(i = 1, 2, \ldots, k)$, and consequently the cost of the survey is proportional to the size of the sample, n_i's are given by

$$n_i = n \frac{p_i S_i}{\sum\limits_{i=1}^{k} p_i S_i} \tag{14}$$

In other words, for a given size of the sample, the allocation in accordance with (14) yields the estimate of the mean with the maximum precision. This result appears to have been first discovered by Tschuprow (1923), but remained unknown until it was rediscovered independently by Neyman (1934). The allocation of the sample in accordance with this formula is, however, known as *Neyman allocation* in literature.

When the population mean is desired to be estimated with a given variance, say V_0, at a minimum cost, we evaluate the constant of proportionality by substituting for n_i from (11) in

$$\sum_{i=1}^{k} \left(\frac{1}{n_i} - \frac{1}{N_i}\right) p_i^2 \mathrm{S}_i^2 = V_0 \tag{15}$$

and obtain, since $p_i = N_i/N$,

$$\frac{1}{\sqrt{\mu}} = \frac{\sum\limits_{i=1}^{k} p_i \mathrm{S}_i \sqrt{c_i}}{V_0 + \frac{1}{N}\sum\limits_{i=1}^{k} p_i \mathrm{S}_i^2} \tag{16}$$

Hence

$$n_i = \frac{p_i \mathrm{S}_i}{\sqrt{c_i}} \cdot \frac{\sum\limits_{i=1}^{k} p_i \mathrm{S}_i \sqrt{c_i}}{V_0 + \frac{1}{N}\sum\limits_{i=1}^{k} p_i \mathrm{S}_i^2} \tag{17}$$

When $c_i = c$, (17) reduces to

$$n_i = p_i \mathrm{S}_i \frac{\sum\limits_{i=1}^{k} p_i \mathrm{S}_i}{V_0 + \frac{1}{N}\sum\limits_{i=1}^{k} p_i \mathrm{S}_i^2} \tag{18}$$

so that the minimum sample required for estimating the mean with fixed variance V_0 is given by

$$n = \frac{\left(\sum\limits_{i=1}^{k} p_i \mathrm{S}_i\right)^2}{V_0 + \frac{1}{N}\sum\limits_{i=1}^{k} p_i \mathrm{S}_i^2} \tag{19}$$

3*a*.4 Size of Sample for Estimating the Mean with a Given Variance under (*a*) Optimum (Neyman), and (*b*) Proportional Allocations

We have seen that for n_i's arbitrary, the variance of the mean is given by

$$V(\bar{y}_w) = \sum_{i=1}^{k} \left(\frac{1}{n_i} - \frac{1}{N_i}\right) p_i^2 \mathrm{S}_i^2 \tag{20}$$

Under Neyman allocation we have

$$n_i = \frac{np_i\mathrm{S}_i}{\sum\limits_{i=1}^{k} p_i\mathrm{S}_i} \tag{21}$$

Substituting for n_i from (21) in (20), we get

$$\begin{aligned} V(\bar{y}_w)_N &= \sum_{i=1}^{k} \left\{ \frac{\sum\limits_{i=1}^{k} p_i\mathrm{S}_i}{np_i\mathrm{S}_i} - \frac{1}{N_i} \right\} p_i^2\mathrm{S}_i^2 \\ &= \frac{1}{n}\left(\sum_{i=1}^{k} p_i\mathrm{S}_i\right)^2 - \frac{1}{N}\sum_{i=1}^{k} p_i\mathrm{S}_i^2 \end{aligned} \tag{22}$$

where the subscript N symbolises the variance of the mean under Neyman allocation.

For proportional allocation of the sample among strata

$$n_i = np_i \tag{23}$$

Substituting in (20), we get

$$\begin{aligned} V(\bar{y}_w)_P &= \sum_{i=1}^{k} \left(\frac{1}{np_i} - \frac{1}{Np_i}\right) p_i^2\mathrm{S}_i^2 \\ &= \left(\frac{1}{n} - \frac{1}{N}\right) \sum_{i=1}^{k} p_i\mathrm{S}_i^2 \\ &= \frac{N-n}{Nn} \sum_{i=1}^{k} p_i\mathrm{S}_i^2 \end{aligned} \tag{24}$$

where the subscript P symbolises the variance under the proportional system of allocation.

It follows from (22) that for estimating the mean from a stratified sample under Neyman allocation with a given variance V_0, the size of sample required is given by

$$n = \frac{\left(\sum\limits_{i=1}^{k} p_i\mathrm{S}_i\right)^2}{V_0 + \frac{1}{N}\sum\limits_{i=1}^{k} p_i\mathrm{S}_i^2} \tag{25}$$

and that under proportional allocation, n is obtained from (24), being given by

$$n = \frac{\sum_{i=1}^{k} p_i \mathrm{S}_i^2}{V_0 + \frac{1}{N} \sum_{i=1}^{k} p_i \mathrm{S}_i^2} \tag{26}$$

Equation (25) is seen to be identical with (19), as is to be expected.

3*a*.5 Comparison of Stratified with Unstratified Simple Random Sampling

We shall discuss the efficiency of stratified sampling first under proportional allocation, then under Neyman allocation and finally under an arbitrary allocation.

(*a*) *Proportional Allocation*

We have seen that

$$V(\bar{y}_w)_P = \left(\frac{1}{n} - \frac{1}{N}\right) \sum_{i=1}^{k} p_i \mathrm{S}_i^2 \tag{27}$$

The variance of the mean under unstratified simple random sampling may be written as

$$V(\bar{y}_n)_{US} = \left(\frac{1}{n} - \frac{1}{N}\right) \mathrm{S}^2 \tag{28}$$

For purposes of comparing (27) with (28), it is necessary to express S^2 in terms of S_i^2.

Now the total sum of squares in the population can be split up into two parts, *viz.*, (*a*) within strata, and (*b*) between strata, in accordance with the identity:

$$\sum_{i=1}^{k} \sum_{j=1}^{N_i} (y_{ij} - \bar{y}_N)^2 \equiv \sum_{i=1}^{k} \sum_{j=1}^{N_i} (y_{ij} - \bar{y}_{N_i} + \bar{y}_{N_i} - \bar{y}_N)^2$$

$$\equiv \sum_{i=1}^{k} \sum_{j=1}^{N_i} (y_{ij} - \bar{y}_{N_i})^2 + \sum_{i=1}^{k} N_i (\bar{y}_{N_i} - \bar{y}_N)^2$$

This can be written as

$$(N-1)\,\mathrm{S}^2 = \sum_{i=1}^{k} (N_i - 1)\,\mathrm{S}_i^2 + \sum_{i=1}^{k} N_i (\bar{y}_{N_i} - \bar{y}_N)^2 \tag{29}$$

For purposes of simplification we will assume N_i to be large enough to permit the approximations

$$\frac{N_i - 1}{N_i} \cong 1 \qquad \text{and} \qquad \frac{N-1}{N} \cong 1$$

On dividing by N, and making use of these approximations, we get

$$\mathbf{S}^2 \cong \sum_{i=1}^{k} p_i \mathbf{S}_i^2 + \sum_{i=1}^{k} p_i (\bar{y}_{N_i} - \bar{y}_N)^2$$

or

$$\sum_{i=1}^{k} p_i \mathbf{S}_i^2 \cong \mathbf{S}^2 - \sum_{i=1}^{k} p_i (\bar{y}_{N_i} - \bar{y}_N)^2 \tag{30}$$

Substituting the result in (27), we have

$$V(\bar{y}_w)_P \cong \left(\frac{1}{n} - \frac{1}{N}\right) \left\{\mathbf{S}^2 - \sum_{i=1}^{k} p_i (\bar{y}_{N_i} - \bar{y}_N)^2\right\} \tag{31}$$

Hence, subtracting (31) from (28), we obtain

$$V(\bar{y}_n)_{US} - V(\bar{y}_w)_P \cong \frac{N-n}{Nn} \sum_{i=1}^{k} p_i (\bar{y}_{N_i} - \bar{y}_N)^2 \tag{32}$$

The expression shows that the more the strata differ in their means, the larger is the gain in precision due to proportional sampling over unstratified simple random sampling.

(*b*) *Neyman Allocation*

Here, again, we shall first express the variance of the mean under Neyman allocation in a form suitable for comparison with the variance in unstratified simple random sampling. We shall make usc of the identity:

$$\sum_{i=1}^{k} p_i (\mathbf{S}_i - \bar{\mathbf{S}}_w)^2 \equiv \sum_{i=1}^{k} p_i \mathbf{S}_i^2 - \left(\sum_{i=1}^{k} p_i \mathbf{S}_i\right)^2 \tag{33}$$

where

$$\bar{\mathbf{S}}_w = \sum_{i=1}^{k} p_i \mathbf{S}_i$$

On substituting for $\left(\sum_{i=1}^{k} p_i \mathrm{S}_i\right)^2$ from (33) in the expression for the variance of the mean under Neyman allocation given in (22), we obtain

$$V(\bar{y}_w)_N = \frac{1}{n}\left[\sum_{i=1}^{k} p_i \mathrm{S}_i^2 - \sum_{i=1}^{k} p_i (\mathrm{S}_i - \bar{\mathrm{S}}_w)^2\right] - \frac{1}{N}\sum_{i=1}^{k} p_i \mathrm{S}_i^2$$

$$= \frac{N-n}{Nn}\sum_{i=1}^{k} p_i \mathrm{S}_i^2 - \frac{1}{n}\sum_{i=1}^{k} p_i (\mathrm{S}_i - \bar{\mathrm{S}}_w)^2 \qquad (34)$$

The first term on the right-hand side of the above equation represents the variance under proportional allocation. Hence

$$V(\bar{y}_w)_P - V(\bar{y}_w)_N = \frac{1}{n}\sum_{i=1}^{k} p_i (\mathrm{S}_i - \bar{\mathrm{S}}_w)^2 \qquad (35)$$

We see that the larger the differences between the strata standard deviations, the larger is the gain in precision of optimum over proportional allocation. Further, on substituting for $V(\bar{y}_w)_P$ from (31), we obtain

$$V(\bar{y}_w)_N \cong \frac{N-n}{Nn}\left\{\mathrm{S}^2 - \sum_{i=1}^{k} p_i (\bar{y}_{N_i} - \bar{y}_N)^2 - \frac{N}{N-n}\sum_{i=1}^{k} p_i (\mathrm{S}_i - \bar{\mathrm{S}}_w)^2\right\} \qquad (36)$$

Since the first term on the right-hand side of (36) represents the variance of the mean of an unstratified sample of n, we may write

$$V(\bar{y}_n)_{US} - V(\bar{y}_w)_N \cong \frac{N-n}{Nn}\left\{\sum_{i=1}^{k} p_i (\bar{y}_{N_i} - \bar{y}_N)^2 + \frac{N}{N-n}\sum_{i=1}^{k} p_i (\mathrm{S}_i - \bar{\mathrm{S}}_w)^2\right\} \qquad (37)$$

The equation (37) shows that the gain in precision of Neyman allocation over unstratified simple random sampling arises from two factors, *viz.*, (*a*) differences between strata means, and (*b*) differences between strata standard deviations.

The above result is exceedingly helpful in devising a scheme of stratification. It suggests that for efficient stratification, the population should be so divided that the differences between the strata means and standard deviations are as large as possible. This is best done in practice by grouping together like units of the population. Thus, geographically contiguous units are usually more alike than those further apart. Consequently, the adoption of geographical proximity as the basis for stratification is expected to lead to a gain in precision apart from being convenient for purposes of organization of field work. On the other hand, experience shows that the gains made from geographical stratification are generally moderate. Another method of setting up strata is to use the information on some correlated character as the basis for stratification. For example, the size of farm is known to be correlated with a number of farm characters like the area under principal crops or the number of livestock; stratification by size of farm in agricultural surveys is, therefore, expected to lead to substantial gains in precision. Example 3.1 at the end of Section 3*a*.10 will serve to illustrate the magnitude of gains recorded from this type of stratification.

(*c*) *Arbitrary Allocation*

When a sample is divided arbitrarily among the strata, the variance of the estimated mean is given by the expression (7); whereas, when it is selected as an unstratified simple random sample, the variance is given by

$$V(\bar{y}_n)_{US} = \left(\frac{1}{n} - \frac{1}{N}\right) S^2$$

We therefore have

$$V(\bar{y}_n)_{US} - V(\bar{y}_w)_S = \left(\frac{1}{n} - \frac{1}{N}\right) S^2 - \sum_{i=1}^{k} \left(\frac{1}{n_i} - \frac{1}{N_i}\right) p_i^2 S_i^2 \tag{38}$$

Substituting for S² from (30), we write

$$V(\bar{y}_n)_{US} - V(\bar{y}_w)_S \cong \left(\frac{1}{n} - \frac{1}{N}\right) \left\{\sum_{i=1}^{k} p_i S_i^2 + \sum_{i=1}^{k} p_i (\bar{y}_{N_i} - \bar{y}_N)^2\right\}$$
$$- \sum_{i=1}^{k} \left(\frac{1}{n_i} - \frac{1}{N_i}\right) p_i^2 S_i^2$$

$$= \sum_{i=1}^{k} \left(\frac{1}{n} - \frac{p_i}{n_i}\right) p_i S_i^2$$

$$+ \left(\frac{1}{n} - \frac{1}{N}\right) \sum_{i=1}^{k} p_i \, (\bar{y}_{N_i} - \bar{y}_N)^2 \qquad (39)$$

The second term on the right-hand side in (39) is always positive, but the first may be positive, zero or negative, depending upon the values of n_i. It is positive when the sample is allocated according to Neyman allocation and we reach the result (37). It is zero when

either

$$n_i = np_i \qquad (40)$$

or

$$n_i = n \frac{p_i S_i^2}{\sum\limits_{i=1}^{k} p_i S_i^2} \qquad (41)$$

giving us (32). As the allocation departs from (40) or (41), the first term may not only become negative but be larger in magnitude than the second, thus making a stratified sample less efficient than an unstratified sample. The result is important and suggests the need for care in the allocation of the sample among the strata.

3*a*.6* Practical Difficulties in Adopting the Neyman Method of Allocation

There are certain limitations to the use of the Neyman allocation in practice which will now be pointed out. If more than one character is to be estimated from a sample survey, then the allocation of the sample into different strata on the basis of any one character, using the Neyman method, may lead to loss in precision on other characters as compared to the method of proportional allocation. If, however, the characters are correlated, or if certain characters are more important than others, then gains in precision on the estimates of the more important characters can still be secured by using the Neyman method of allocation. However, the more severe limitation on the use of the Neyman allocation is the absence of the knowledge of S_i's. One method of overcoming this limitation is to estimate S_i's from a preliminary sample of n' (Sukhatme, 1935). These estimates will, however, be subject to standard errors and it is, therefore, possible that we would be worse off than if we had used the method of proportional allocation. The problems to be considered are then, (*a*) how much would the variance increase, on the average, if the allocation is based on estimated values of $\mathbf{S}_i$, (*b*) how does it compare with the variance from proportional allocation, and (*c*) how large should be the size of the preliminary sample in order that Neyman allocation may give a more precise estimate than proportional allocation ?

Let s_i represent an unbiased estimate of $\mathbf{S}_i/v$, based on a sample of size n', v denoting a constant, so that

$$E\,(vs_i) = \mathbf{S}_i \qquad (i = 1, 2, \ldots, k) \tag{42}$$

The allocation of the total sample among the different strata will now be made in accordance with the formula*

$$n_i = \frac{np_i s_i}{\sum\limits_{i=1}^{k} p_i s_i} \tag{43}$$

Substituting in (7), we obtain

$$\begin{aligned} V\left(\bar{y}_w \left| \begin{matrix} s_1 \\ \vdots \\ s_k \end{matrix} \right.\right)_N &= \sum_{i=1}^{k} \left\{ \frac{\sum\limits_{i=1}^{k} p_i s_i}{np_i s_i} - \frac{1}{Np_i} \right\} p_i^2 \mathbf{S}_i^2 \\ &= \frac{1}{n}\left(\sum_{i=1}^{k} p_i s_i\right)\left(\sum_{i=1}^{k} p_i \frac{\mathbf{S}_i^2}{s_i}\right) - \frac{1}{N}\sum_{i=1}^{k} p_i \mathbf{S}_i^2 \\ &= \frac{1}{n}\left\{\sum_{i=1}^{k} p_i^2 \mathbf{S}_i^2 + \sum_{i \neq j=1}^{k} p_i p_j \mathbf{S}_j^2 \frac{s_i}{s_j}\right\} - \frac{1}{N}\sum_{i=1}^{k} p_i \mathbf{S}_i^2 \end{aligned} \tag{44}$$

This expression involves s_i and we are consequently not in a position to say whether this will give a smaller value or not as compared to that for proportional allocation, *viz.*,

$$V\,(\bar{y}_w)_P = \left(\frac{1}{n} - \frac{1}{N}\right) \sum_{i=1}^{k} p_i \mathbf{S}_i^2 \tag{45}$$

* Where, as in this case, the decision regarding the size of the additional sample to be drawn from each stratum depends upon the results of the first sample, the procedure is essentially what is called sequential sampling.

We might, however, obtain the average value of (44) in samples of n' and examine how it compares with (45).

Now it can be shown (Sukhatme, 1935) that if the variate under study can be considered as normally distributed and consequently s_i is distributed as χ, the average value of the right-hand side in (44) is given approximately by

$$\frac{1}{n}\left\{\sum_{i=1}^{k} p_i^2\mathbf{S}_i^2 + \theta \sum_{i\neq j=1}^{k} p_i p_j \mathbf{S}_i\mathbf{S}_j\right\} - \frac{1}{N}\sum_{i=1}^{k} p_i\mathbf{S}_i^2 \tag{46}$$

where

$$\theta = 1 + \frac{1}{2n'} \tag{47}$$

Substituting for θ from (47) in (46), we obtain

$$\begin{aligned} E\left\{V(\bar{y}_w) \,\middle|\, \begin{matrix} n' \\ s_i \end{matrix}\right\}_N &= \frac{1}{n}\left(\sum_{i=1}^{k} p_i\mathbf{S}_i\right)^2 - \frac{1}{N}\sum_{i=1}^{k} p_i\mathbf{S}_i^2 \\ &\quad + \frac{1}{2nn'}\left\{\left(\sum_{i=1}^{k} p_i\mathbf{S}_i\right)^2 - \sum_{i=1}^{k} p_i^2\mathbf{S}_i^2\right\} \\ &= \left(\frac{1}{n} - \frac{1}{N}\right)\sum_{i=1}^{k} p_i\mathbf{S}_i^2 - \frac{1}{n}\sum_{i=1}^{k} p_i\,(\mathbf{S}_i - \bar{\mathbf{S}}_w)^2 \\ &\quad + \frac{1}{2nn'}\left\{\left(\sum_{i=1}^{k} p_i\mathbf{S}_i\right)^2 - \sum_{i=1}^{k} p_i^2\mathbf{S}_i^2\right\} \end{aligned} \tag{48}$$

which, on using (30), can also be written as

$$\begin{aligned} E\left\{V(\bar{y}_w) \,\middle|\, \begin{matrix} n' \\ s_i \end{matrix}\right\}_N &\cong \frac{N-n}{Nn}\left\{\mathbf{S}^2 - \sum_{i=1}^{k} p_i\,(\bar{y}_{N_i} - \bar{y}_N)^2\right. \\ &\qquad \left. - \frac{N}{N-n}\sum_{i=1}^{k} p_i\,(\mathbf{S}_i - \bar{\mathbf{S}}_w)^2\right\} \\ &\quad + \frac{1}{2nn'}\left\{\left(\sum_{i=1}^{k} p_i\mathbf{S}_i\right)^2 - \sum_{i=1}^{k} p_i^2\mathbf{S}_i^2\right\} \end{aligned} \tag{49}$$

The first part on the right-hand side denotes the variance of the mean under Neyman allocation when the $\mathbf{S}_i$ are known. Consequently, when the $\mathbf{S}_i$ are estimated from a preliminary sample of size n', this variance is seen to increase, on an average, by

$$\frac{1}{2nn'}\left\{\left(\sum_{i=1}^{k} p_i\mathbf{S}_i\right)^2 - \sum_{i=1}^{k} p_i^2\mathbf{S}_i^2\right\} \tag{50}$$

Comparing (48) with the value of the variance under proportional allocation given in (45), we notice that the condition that the Neyman allocation may not, on the average, lead to loss of precision as compared to proportional allocation is

$$\frac{1}{2n'}\left\{\left(\sum_{i=1}^{k} p_i\mathbf{S}_i\right)^2 - \sum_{i=1}^{k} p_i^2\mathbf{S}_i^2\right\} \leqq \sum_{i=1}^{k} p_i\,(\mathbf{S}_i - \bar{\mathbf{S}}_w)^2 \tag{51}$$

or

$$n' \geqq \frac{\left(\sum\limits_{i=1}^{k} p_i\mathbf{S}_i\right)^2 - \sum\limits_{i=1}^{k} p_i^2\mathbf{S}_i^2}{2\sum\limits_{i=1}^{k} p_i\,(\mathbf{S}_i - \bar{\mathbf{S}}_w)^2} \tag{52}$$

The above result can be derived more simply by an alternative method given by Evans (1951).

Let, as before,

$$vs_i = \mathbf{S}_i + \epsilon_i$$

where

$$E\,(\epsilon_i) = 0, \qquad E\,(\epsilon_i^2) = v^2V\,(s_i)$$

and

$$vs_j = \mathbf{S}_j + \epsilon_j$$

where

$$E\,(\epsilon_j) = 0, \qquad E\,(\epsilon_j^2) = v^2V\,(s_j)$$

Then s_i/s_j can be expressed as

$$\begin{aligned}\frac{s_i}{s_j} &= \frac{\mathbf{S}_i}{\mathbf{S}_j}\left(1 + \frac{\epsilon_i}{\mathbf{S}_i}\right)\left(1 + \frac{\epsilon_j}{\mathbf{S}_j}\right)^{-1}\\ &= \frac{\mathbf{S}_i}{\mathbf{S}_j}\left(1 + \frac{\epsilon_i}{\mathbf{S}_i}\right)\left(1 - \frac{\epsilon_j}{\mathbf{S}_j} + \frac{\epsilon_j^2}{\mathbf{S}_j^2} - \cdots\right)\end{aligned}$$

On expanding the right-hand side and neglecting terms involving powers of ϵ higher than the second and taking expectations, we obtain

$$E\left(\frac{s_i}{s_j}\right) \cong \frac{\mathbf{S}_i}{\mathbf{S}_j}\,(1 + C_j^2)$$

where C_j is the coefficient of variation of s_j given by $v\,\sqrt{V(\overline{s_j})}/\mathbf{S}_j$. Assuming $C_j = C$ $(j = 1, 2, \ldots, k)$, we obtain

$$E\left(\frac{s_i}{s_j}\right) \cong \frac{\mathbf{S}_i}{\mathbf{S}_j}\,(1 + C^2) \tag{53}$$

Taking expectations of both sides in (44) and substituting from (53), we have

$$\begin{aligned}
E\left\{V(\bar{y}_w)\,\middle|\,\begin{matrix}n'\\ s_i\end{matrix}\right\}_N &\cong \frac{1}{n}\left\{\sum_{i=1}^{k} p_i^2\mathbf{S}_i^2 + \sum_{i \neq j = 1}^{k} p_i p_j \mathbf{S}_i \mathbf{S}_j\,(1 + C^2)\right\} \\
&\qquad\qquad - \frac{1}{N}\sum_{i=1}^{k} p_i \mathbf{S}_i^2 \\
&= V(\bar{y}_w)_N + \frac{1}{n}\,C^2\left\{\left(\sum_{i=1}^{k} p_i \mathbf{S}_i\right)^2 - \sum_{i=1}^{k} p_i^2 \mathbf{S}_i^2\right\} \\
&= V(\bar{y}_w)_P - \frac{1}{n}\sum_{i=1}^{k} p_i\,(\mathbf{S}_i - \bar{\mathbf{S}}_w)^2 \\
&\qquad + \frac{1}{n}\,C^2\left\{\left(\sum_{i=1}^{k} p_i \mathbf{S}_i\right)^2 - \sum_{i=1}^{k} p_i^2 \mathbf{S}_i^2\right\}
\end{aligned} \tag{54}$$

Clearly, this expression will have a smaller value as compared to that under proportional allocation if the sum of the last two terms is negative, *i.e.*, if

$$C^2 \leq \frac{\sum\limits_{i=1}^{k} p_i\,(\mathbf{S}_i - \bar{\mathbf{S}}_w)^2}{\left(\sum\limits_{i=1}^{k} p_i \mathbf{S}_i\right)^2 - \sum\limits_{i=1}^{k} p_i^2 \mathbf{S}_i^2} \tag{55}$$

From Sections $2a.10$ and $2a.11$ of the last chapter, we know that, for samples of n', C^2 is approximately given by $(\beta_2 - 1)/4n'$, so that the size of the preliminary sample should be such that

$$n' \geqq \frac{\beta_2 - 1}{4} \cdot \frac{\left(\sum\limits_{i=1}^{k} p_i\mathbf{S}_i\right)^2 - \sum\limits_{i=1}^{k} p_i^2\mathbf{S}_i^2}{\sum\limits_{i=1}^{k} p_i\,(\mathbf{S}_i - \bar{\mathbf{S}}_w)^2} \tag{56}$$

in order that the Neyman allocation may give, on the average, a more precise estimate than the method of proportional allocation.

When $\beta_2 = 3$, the value of n' reduces to that given by (52).

It will be seen that the larger the variability among $\mathbf{S}_i$'s, the smaller will be the value of n'. Consequently, unless $\mathbf{S}_i$'s are close to one another, even moderately small preliminary samples will give, on the average, more precise results than proportional allocation. If, however, the size of the preliminary sample is found to be so large as to make the preliminary inquiry not worth while and if the study of several characters is included in the sample, proportional allocation would be preferable.

3*a*.7 Evaluation from the Sample of the Gain in Precision due to Stratification

In comparing the precision of the stratified with unstratified simple random sampling in Section $3a.5$, we assumed that the population values of the strata means and standard deviations were known. Usually, however, this will not be the case. What is available is only a stratified sample and the problem is to estimate the gain in precision due to stratification.

Let $n_1, n_2, \ldots, n_k$ represent the stratified sample. Then the variance of the sample mean is

$$V(\bar{y}_w)_S = \sum_{i=1}^{k}\left(\frac{1}{n_i} - \frac{1}{N_i}\right) p_i^2\mathbf{S}_i^2.$$

and its estimate is clearly given by

$$\text{Est. } V(\bar{y}_w)_S = \sum_{i=1}^{k}\left(\frac{1}{n_i} - \frac{1}{N_i}\right) p_i^2 s_i^2 \tag{57}$$

where s_i^2 is the mean square in the sample drawn from the i-th stratum. If the total sample had been selected by the procedure of simple random sampling without stratification, then the variance of the sample mean would be

$$V(\bar{y}_n)_{US} = \frac{N-n}{N} \cdot \frac{\mathbf{S}^2}{n} \tag{58}$$

Its estimate cannot, however, be obtained by substituting the mean square for the total sample in place of $\mathbf{S}^2$. For, although within each stratum the sample is selected by the method of simple random sampling, the total sample cannot be considered to have been so selected from the population as a whole. The problem, therefore, is to estimate $\mathbf{S}^2$, given

$$\bar{y}_{n_1}, \bar{y}_{n_2}, \ldots, \bar{y}_{nk} \qquad \text{and} \qquad s_1^2, s_2^2, \ldots, s_k^2$$

We have, from (29),

$$\mathbf{S}^2 = \frac{1}{N-1} \sum_{i=1}^{k} (N_i - 1)\, \mathbf{S}_i^2 + \frac{N}{N-1} \sum_{i=1}^{k} p_i (\bar{y}_{Ni} - \bar{y}_N)^2 \tag{59}$$

Since s_i^2 provides an unbiased estimate of $\mathbf{S}_i^2$, the problem of estimating $\mathbf{S}^2$ reduces to the estimation of

$$\sum_{i=1}^{k} p_i (\bar{y}_{N_i} - \bar{y}_N)^2 \tag{60}$$

from the sample.

Let

$$\bar{y}_{n_i} = \bar{y}_{N_i} + \epsilon_i \tag{61}$$

where

$$E(\epsilon_i) = 0, \qquad \text{and} \qquad E(\epsilon_i^2) = \frac{N_i - n_i}{N_i n_i} \mathbf{S}_i^2$$

On squaring both sides of (61), we have

$$\bar{y}_{n_i}^2 = \bar{y}_{N_i}^2 + \epsilon_i^2 + 2\bar{y}_{N_i}\epsilon_i \tag{62}$$

Taking expectations, we obtain

$$E(\bar{y}_{n_i}^2) = \bar{y}_{N_i}^2 + \frac{N_i - n_i}{N_i n_i} \mathbf{S}_i^2 \tag{63}$$

$$\text{Est. } \{V(\bar{y}_n)_{US} - V(\bar{y}_w)_S\} = \frac{1}{n} \sum_{i=1}^{k} p_i s_i^2 - \sum_{i=1}^{k} \frac{p_i^2 s_i^2}{n_i} + \frac{N-n}{(N-1)\,n} \left\{ \sum_{i=1}^{k} p_i (\bar{y}_{n_i} - \bar{y}_w)^2 - \sum_{i=1}^{k} p_i (1 - p_i) \frac{s_i^2}{n_i} \right\} \qquad (73)$$

Taking $N/(N-1) \cong 1$, we obtain

$$\text{Est. } \{V(\bar{y}_n)_{US} - V(\bar{y}_w)_S\} \cong \frac{1}{n} \sum_{i=1}^{k} p_i s_i^2 - \sum_{i=1}^{k} \frac{p_i^2 s_i^2}{n_i} + \frac{N-n}{Nn} \times \left\{ \sum_{i=1}^{k} p_i (\bar{y}_{n_i} - \bar{y}_w)^2 - \sum_{i=1}^{k} p_i (1 - p_i) \frac{s_i^2}{n_i} \right\} \qquad (74)$$

The ratio of (73) or (74) to (57), expressed as a percentage, gives the estimate of the gain in efficiency due to stratification.

These results assume a particularly simple form in the case of proportional sampling for which $\bar{y}_w = \bar{y}_n$. Equation (57) then becomes

$$\text{Est. } V(\bar{y}_w)_P = \frac{N-n}{Nn} \sum_{i=1}^{k} p_i s_i^2 \qquad (75)$$

and the first two terms in (74) vanish. The net reduction in variance due to stratification is, therefore, given by the last two terms in (74). On substituting n_i/n for p_i, this takes the value

$$\text{Est. } V(\bar{y}_n)_{US} - \text{Est. } V(\bar{y}_w)_P = \frac{N-n}{Nn} \times \left\{ \frac{1}{n} \sum_{i=1}^{k} n_i (\bar{y}_{n_i} - \bar{y}_w)^2 - \frac{1}{n} \sum_{i=1}^{k} \left(1 - \frac{n_i}{n}\right) s_i^2 \right\} \qquad (76)$$

When the finite multiplier is assumed to be unity and S_i^2 is constant, say $S_i^2 = S_w^2$ $(i = 1, 2, \ldots, k)$, the best unbiased estimate of the latter is obtained by pooling the sum of squares within strata for the sample. We write

$$\text{Est. } \mathbf{S}_w^2 = \frac{1}{n-k} \sum_{i=1}^{k} \sum_{j}^{n_i} (y_{ij} - \bar{y}_{n_i})^2 = s_w^2, \text{ say} \tag{77}$$

since

$$E\,(s_w^2) = \frac{1}{n-k}\, E \left\{ \sum_{i=1}^{k} \sum_{j}^{n_i} (y_{ij} - \bar{y}_{n_i})^2 \right\}$$

$$= \frac{1}{n-k}\, E \sum_{i=1}^{k} E \sum_{j}^{n_i} (y_{ij} - \bar{y}_{n_i})^2$$

$$= \frac{1}{n-k}\, E \sum_{i=1}^{k} (n_i - 1)\, \mathbf{S}_i^2$$

$$= \mathbf{S}_w^2$$

Let

$$\sum_{i=1}^{k} n_i\,(\bar{y}_{n_i} - \bar{y}_w)^2 = (k-1)\,\bar{n} s_b^2 \tag{78}$$

where $\bar{n} = n/k$. Substituting from (77) and (78) in (72), we have

$$\text{Est. } V\,(\bar{y}_n)_{US} \cong \frac{1}{n^2}\, \{(n-k+1)\, s_w^2 + (k-1)\, \bar{n} s_b^2\} \tag{79}$$

The quantities s_w^2 and $\bar{n}s_b^2$ are called the mean squares within and between strata respectively, and are best calculated from what is familiarly known as the analysis of variance table given below:

Source of Variation	D.F.	Sum of Squares	Mean Square
Between Strata ..	$k-1$	$\sum_{i=1}^{k} n_i\,(\bar{y}_{n_i} - \bar{y}_n)^2$	$\bar{n}s_b^2$
Within Strata ..	$n-k$	$\sum_{i=1}^{k} \sum_{j}^{n_i} (y_{ij} - \bar{y}_{n_i})^2$	s_w^2
Total ..	$n-1$	$\sum_{i=1}^{k} \sum_{j}^{n_i} (y_{ij} - \bar{y}_n)^2$	s^2

Also, from (75),

$$\text{Est. } V(\bar{y}_w)_P \cong \frac{s_w^2}{n} \tag{80}$$

An estimate of the reduction in variance is now given by subtracting (80) from (79) or directly from (76), and equals

$$\text{Est. } V(\bar{y}_n)_{US} - \text{Est. } V(\bar{y}_w)_P = \frac{k-1}{n^2}\{\bar{n}s_b^2 - s_w^2\} \tag{81}$$

The ratio of (81) to (80) gives the relative gain in precision due to stratification and equals

$$\frac{k-1}{n}\left\{\frac{\bar{n}s_b^2}{s_w^2} - 1\right\} \tag{82}$$

The efficiency of stratification is sometimes calculated directly by comparing the overall mean square s^2 with s_w^2, the relative gain in precision being given by

$$\begin{aligned}\frac{s^2}{s_w^2} - 1 &= \frac{(n-k)\,s_w^2 + (k-1)\,\bar{n}s_b^2}{(n-1)\,s_w^2} - 1\\ &= \frac{k-1}{n-1}\left(\frac{\bar{n}s_b^2}{s_w^2} - 1\right)\end{aligned} \tag{83}$$

The gain in precision estimated this way is $n/(n-1)$ times the value in (82), which is not likely to be of material difference in large samples, provided the sample is allocated in proportion to the sizes of the different strata. When the sample is not so allocated, neither (82) nor (83) is likely to be satisfactory. The exact expression given by the ratio of (73) to (57) should be used in that case.

3*a*.8 Use of Strata Sizes for Improving the Precision of an Unstratified Sample

Stratified sampling presupposes the knowledge of the strata sizes as well as the availability of the lists of sampling units for the different strata. The latter are not, however, always available. Thus, the classification of a population by age is known from the census tables although the lists of persons belonging to different age groups may not be available for the selection of samples from the different age groups. Consequently, it is not possible to know in advance to which stratum a sampling unit belongs until it is contacted in the course of the survey itself. While the sample in such cases has necessarily to be selected by the method of unstratified random sampling, we can always classify the selected sample by the strata and treat it as if it were a stratified sample. In this section we shall examine the gain in precision arising from such a treatment.

If the sample is to be treated as if it were a stratified sample, then $\bar{y}_w$ would be the appropriate estimate of the population mean. This is easily seen to be an unbiased estimate of the population mean, since

$$\begin{aligned} E(\bar{y}_{n_i}) &= E\{E(\bar{y}_{n_i} \mid n_i)\} \\ &= E(\bar{y}_{N_i}) \\ &= \bar{y}_{N_i} \end{aligned}$$

Hence

$$E(\bar{y}_w) = \bar{y}_N \tag{84}$$

For fixed $n_1, n_2, \ldots, n_k$, the variance of $\bar{y}_w$ is given by

$$V(\bar{y}_w) = \sum_{i=1}^{k} \left(\frac{1}{n_i} - \frac{1}{N_i}\right) p_i^2 \mathrm{S}_i^2 \tag{85}$$

However n_i $(i = 1, 2, \ldots, k)$, varies from sample to sample. Consequently, (85) is not directly comparable with the variance of an unstratified sample. We can, however, examine how this method compares, on an average, with that of an unstratified sample. An exact expression for the average value of (85) cannot be obtained. For large values of n, however, and for large N we may use the result (112) of Section 2*a*.19 of the last Chapter and write

$$E\left(\frac{1}{n_i}\right) \cong \frac{1}{np_i} + \frac{1-p_i}{n^2 p_i^2} + O\left(\frac{1}{n^4}\right) \tag{86}$$

Taking expectations in (85) and using (86), we have

$$\begin{aligned} E\{V(\bar{y}_w)_{US}\} &\cong \sum_{i=1}^{k} \left\{\frac{1}{np_i} + \frac{1-p_i}{n^2 p_i^2} + O\left(\frac{1}{n^4}\right) - \frac{1}{Np_i}\right\} p_i^2 \mathrm{S}_i^2 \\ &= \frac{1}{n}\left\{\left(1 - \frac{n}{N} - \frac{1}{n}\right) \sum_{i=1}^{k} p_i \mathrm{S}_i^2 + \frac{1}{n} \sum_{i=1}^{k} \mathrm{S}_i^2\right\} \\ &\qquad + O\left(\frac{1}{n^4}\right) \end{aligned} \tag{87}$$

and to the first approximation

$$E\{V(\bar{y}_w)_{US}\} \cong \frac{N-n}{Nn} \sum_{i=1}^{k} p_i \mathbf{S}_i^2 + \frac{1}{n^2} \sum_{i=1}^{k} (1-p_i)\, \mathbf{S}_i^2 \qquad (88)$$

It is seen that the first term in (88) is the variance of the mean of a stratified sample with proportional allocation. We, therefore, see that the adjustment of the results of an unstratified random sample as if it were stratified gives almost as high precision as a stratified sample with proportional allocation, provided the sample is large. The result is obvious otherwise also, for, a large sample is expected to be distributed in proportion to the sizes of the strata.

3*a*.9 Effect of Increasing the Number of Strata on the Precision of the Estimate

The variance of the estimate of the population mean from a stratified sample depends upon

(i) the strata values of p_i and $\mathbf{S}_i$, and

(ii) the sample numbers n_i.

We shall assume that n_i is proportional to p_i, so that the variance will now depend only on the strata values of p_i and $\mathbf{S}_i$, being given by

$$V(\bar{y}_w)_P = \left(\frac{1}{n} - \frac{1}{N}\right) \sum_{i=1}^{k} p_i \mathbf{S}_i^2 \qquad (89)$$

The smaller the strata the more alike will presumably be the sampling units comprising them and the smaller, therefore, will be the values of $\mathbf{S}_i^2$. We may, therefore, expect that under proportional allocation the precision of the estimate will generally increase as the number of strata increases.

For small departures from proportionality, the effect of increasing the number of strata is best studied with the help of (88). The first term in this equation, it will be noticed, is identical with (89) and will presumably decrease as k increases. On the other hand, the contribution of the second term to the variance of $\bar{y}_w$

will increase as k increases. For N large and S_i^2 equal to, say S_w^2, (88) may be written as

$$E\{V(\bar{y}_w)_{US}\} \cong \frac{1}{n}\left\{S_w^2 + \frac{k-1}{n} S_w^2\right\} \tag{90}$$

Now, S_w^2 will ordinarily decrease as k increases, but $(k-1)\,S_w^2$ will increase as k increases, at a rate ordinarily greater than the rate of decrease in S_w^2. We conclude therefore that, for given n, a stage in the value of k may be reached beyond which stratification may not add to the precision of the estimate.

3*a*.10 Effects of Inaccuracies in Strata Sizes

It sometimes happens that the knowledge of strata sizes though available is not exact, as, for example, when it is based on old census data which are out of date for use in current surveys or when derived from a sample. Thus, in a study of farm facts, the survey may be organized in two stages, first selecting a large sample for estimating the strata sizes and then a second sample out of the first for the main purpose of the survey. This latter procedure is called *double sampling* (Neyman, 1938). In this section, we shall examine the effect of inaccuracies in strata sizes on the estimate of the mean and its variance.

Case I. Strata Sizes Fixed

Let p_i denote the true but unknown weight of the i-th stratum and p_i' the inaccurate weight which is known. The sample estimate of the population mean is then given by

$$\text{Est. } \bar{y}_N = \sum_{i=1}^{k} p_i' \bar{y}_{n_i} \tag{91}$$

For fixed p_i''s, this is a biased estimate of the population mean, for, in general

$$E\left(\sum_{i=1}^{k} p_i' \bar{y}_{n_i} \,\middle|\, \begin{matrix} p_1' \\ \vdots \\ p_k' \end{matrix}\right) = \sum_{i=1}^{k} p_i' \bar{y}_{N_i} \neq \sum_{i=1}^{k} p_i \bar{y}_{N_i} \tag{92}$$

showing that the estimate is biased by

$$\sum_{i=1}^{k} (p_i' - p_i)\,\bar{y}_{N_i} \tag{93}$$

To obtain the sampling variance, we write

$$V\left(\sum_{i=1}^{k} p_i'\bar{y}_{n_i}\middle|\begin{matrix}p_1'\\ \vdots\\ p_k'\end{matrix}\right)= E\left\{\left(\sum_{i=1}^{k} p_i'\bar{y}_{n_i} - E\left(\sum_{i=1}^{k} p_i'\bar{y}_{n_i}\right)\right)^2\middle|\begin{matrix}p_1'\\ \vdots\\ p_k'\end{matrix}\right\}$$

$$= E\left\{\left(\sum_{i=1}^{k} p_i'\left(\bar{y}_{n_i} - \bar{y}_{N_i}\right)\right)^2\middle|\begin{matrix}p_1'\\ \vdots\\ p_k'\end{matrix}\right\}$$

$$= E\left\{\sum_{i=1}^{k} p_i'^2\left(\bar{y}_{n_i} - \bar{y}_{N_i}\right)^2\right.$$

$$\left. + \sum_{i\neq j=1}^{k} p_i'p_j'\left(\bar{y}_{n_i} - \bar{y}_{N_i}\right)\left(\bar{y}_{n_j} - \bar{y}_{N_j}\right)\middle|\begin{matrix}p_1'\\ \vdots\\ p_k'\end{matrix}\right\} \tag{94}$$

For fixed p_i''s the second term is clearly zero, and we are left with

$$V\left(\sum_{i=1}^{k} p_i'\bar{y}_{n_i}\middle|\begin{matrix}p_1'\\ \vdots\\ p_k'\end{matrix}\right) = \sum_{i=1}^{k} p_i'^2 E\left(\bar{y}_{n_i} - \bar{y}_{N_i}\right)^2$$

$$= \sum_{i=1}^{k} p_i'^2 \frac{N_i - n_i}{N_i n_i} \mathbf{S}_i^2 \tag{95}$$

The mean square error will be the sum of (95) and the square of the bias term in (93). We have

$$M.S.E.\left(\sum_{i=1}^{k} p_i'\bar{y}_{n_i}\middle|\begin{matrix}p_1'\\ \vdots\\ p_k'\end{matrix}\right) = \sum_{i=1}^{k} p_i'^2 \frac{N_i - n_i}{N_i n_i} \mathbf{S}_i^2$$

$$+ \left\{\sum_{i=1}^{k} \left(p_i' - p_i\right)\bar{y}_{N_i}\right\}^2 \tag{96}$$

If the sample were selected by the method of simple random sampling without stratification, the mean square error of the estimate would be identical with its variance, and would be simply

$$V(\bar{y}_n)_{US} = \frac{N-n}{N}\frac{\mathbf{S}^2}{n} \tag{97}$$

As n increases, (97) will decrease and so will the first term in (96). The bias term is, however, independent of the sample size. It follows that (96) may assume a larger value than (97) beyond a certain n, making stratified sampling less accurate than simple random sampling.

An example will help to illustrate the point. Suppose that according to an agricultural census taken in an earlier year, 80% of the holdings were below 5 acres. Information for the current year is not available but we will assume that the percentage of holdings below 5 acres has increased to 85. Suppose, further, that we have selected a stratified sample of n holdings allocated in proportion to the known sizes of the two strata. Then, clearly, the sample estimate of the population mean of the character under study will be calculated from

$$\bar{y}_w = \cdot 80\, \bar{y}_{n_1} + \cdot 20\, \bar{y}_{n_2} \tag{98}$$

The true population mean will, however, be

$$\bar{y}_N = \cdot 85\, \bar{y}_{N_1} + \cdot 15\, \bar{y}_{N_2} \tag{99}$$

The expected value of the estimate in (98) will be

$$E\,(\bar{y}_w) = \cdot 80\, \bar{y}_{N_1} + \cdot 20\, \bar{y}_{N_2} \tag{100}$$

so that the sample estimate will be biased by

$$E\,(\bar{y}_w) - \bar{y}_N = - \cdot 05\, \bar{y}_{N_1} + \cdot 05\, \bar{y}_{N_2} \tag{101}$$

The mean square error of the sample estimate in (98) will be composed of two parts, as shown in (96). Assuming that the population is large so that finite multipliers can be ignored, this will be given by

$$M.S.E.\,(\bar{y}_w) \cong \frac{1}{n}\,\{\cdot 80\, S_1{}^2 + \cdot 20\, S_2{}^2\} + \{\cdot 05\,(\bar{y}_{N_1} - \bar{y}_{N_2})\}^2 \tag{102}$$

If we had selected the sample by the unstratified method of simple random sampling, the same estimate would be unbiased and its variance given by

$$\frac{S^2}{n} \cong \frac{1}{n}\,\{\cdot 85 S_1{}^2 + \cdot 15 S_2{}^2 + \cdot 85\,(\bar{y}_{N_1} - \bar{y}_N)^2 + \cdot 15\,(\bar{y}_{N_2} - \bar{y}_N)^2\} \tag{103}$$

Now, suppose, the actual values of S_1^2, S_2^2, $\bar{y}_{N_1}$ and $\bar{y}_{N_2}$ are as in the following table:

Stratum	S_i^2	$\bar{y}_{Ni}$
1	1	1
2	1	2

Then,

$$M.S.E.\,(\bar{y}_w) = \frac{1}{n} + \cdot 0025 \tag{104}$$

and

$$V\,(\bar{y}_n)_{US} = \frac{1 \cdot 1275}{n} \tag{105}$$

The table below gives the values of (104) and (105) for five different values of n.

n	$M.S.E.\,(\bar{y}_w)$	$V(\bar{y}_n)_{US}$
25	·0425	·04510
50	·0225	·02255
100	·0125	·01128
200	·0075	·00564
400	·0050	·00282

It will be seen that for small n, the actual mean square error of the stratified sample is smaller than that of the unstratified simple random sample, but the superiority is lost after $n = 51$. With a larger size of sample, the bias assumes still larger proportions. It must, however, be pointed out that the bias will not be known in practice and consequently the variance of the mean of a stratified sample will continue to be estimated by the first term in (96), thereby under-estimating the variance.

Case II. Double Sampling

We shall now consider the case of double sampling where p_i's are estimated from a preliminary simple random sample and the character under study is observed on a sub-sample selected from the preliminary sample. Let Q denote the size of the preliminary sample, Q_i the number of units in the i-th stratum and n_i the size of the sub-sample chosen out of Q_i $(i = 1, 2, \ldots, k)$. We have

$$\text{Est. } p_i = \frac{Q_i}{Q}$$

$$= p_i'$$

and

$$\sum_{i=1}^{k} n_i = n$$

The estimate $\sum_{i=1}^{k} p_i'\bar{y}_{n_i}$ is now clearly the unbiased estimate of $\bar{y}_N$. For, from (92), we have

$$E\left(\sum_{i=1}^{k} p_i'\bar{y}_{n_i}\right) = E\,E\left(\sum_{i=1}^{k} p_i'\bar{y}_{ni} \,\middle|\, \begin{matrix} p_1' \\ \vdots \\ p_k' \end{matrix}\right)$$

$$= E\left(\sum_{i=1}^{k} p_i'\bar{y}_{N_i}\right)$$

$$= \sum_{i=1}^{k} p_i\bar{y}_{N_i} \tag{106}$$

To obtain the variance of the estimate, all that we need do is to obtain the expected value of the terms on the right-hand side in (96) for variation in p_i''s in repeated samples of Q. We write

$$V\left(\sum_{i=1}^{k} p_i'\bar{y}_{n_i}\right) = \sum_{i=1}^{k} E(p_i'^2)\,\frac{N_i - n_i}{N_i n_i}\,S_i^2 + E\left(\sum_{i=1}^{k} (p_i' - p_i)\,\bar{y}_{N_i}\right)^2 \tag{107}$$

On expanding the second term on the right-hand side, we have

$$V\left(\sum_{i=1}^{k} p_i'\bar{y}_{n_i}\right) = \sum_{i=1}^{k} E\,(p_i'^2)\,\frac{N_i - n_i}{N_i n_i}\,S_i^2 + \sum_{i=1}^{k} E\,(p_i' - p_i)^2\,\bar{y}_{N_i}^2 + \sum_{i \neq j = 1}^{k} E\,(p_i' - p_i)\,(p_j' - p_j)\,\bar{y}_{N_i}\bar{y}_{N_j} \tag{108}$$

Now from (71), (78) and (98) of Chapter II, we have

$$E(p_i'^2) = p_i^2 + \frac{N-Q}{N-1}\,\frac{p_i(1-p_i)}{Q} \tag{109}$$

$$E(p_i' - p_i)^2 = \frac{N-Q}{N-1}\,\frac{p_i(1-p_i)}{Q} \tag{110}$$

and

$$E\{(p_i' - p_i)(p_j' - p_j)\} = -\frac{N-Q}{N-1}\,\frac{1}{Q}\,p_i p_j \tag{111}$$

On substituting from (109), (110) and (111) in (108), we have

$$\begin{aligned} V\left(\sum_{i=1}^{k} p_i' \bar{y}_{n_i}\right) = \sum_{i=1}^{k}\left(p_i^2 + \frac{N-Q}{N-1}\,\frac{p_i(1-p_i)}{Q}\right)\frac{N_i - n_i}{N_i n_i}\,\mathrm{S}_i^2 \\ - \sum_{i \neq j = 1}^{k} \frac{N-Q}{N-1}\,\frac{1}{Q}\,p_i p_j \bar{y}_{N_i} \bar{y}_{N_j} \\ + \sum_{i=1}^{k} \frac{N-Q}{N-1}\,\frac{1}{Q}\,p_i(1-p_i)\,\bar{y}_{N_i}^2 \end{aligned} \tag{112}$$

On combining the last two terms, we obtain

$$\begin{aligned} V\left(\sum_{i=1}^{k} p_i' \bar{y}_{n_i}\right) = \sum_{i=1}^{k}\left(p_i^2 + \frac{N-Q}{N-1}\,\frac{p_i(1-p_i)}{Q}\right)\frac{N_i - n_i}{N_i n_i}\,\mathrm{S}_i^2 \\ + \frac{N-Q}{N-1}\,\frac{1}{Q}\sum_{i=1}^{k} p_i(\bar{y}_{N_i} - \bar{y}_N)^2 \end{aligned} \tag{113}$$

The term involving the differences between the strata means on the right-hand side of (113) can also be written directly from (107). For, the second term in (107) clearly represents the variance of the mean of a simple random sample of size Q drawn from a population of size N divided into k classes with all the N_i values in the i-th class equal to $\bar{y}_{N_i}$ each, $(i = 1, 2, \ldots, k)$, so that

$$E\left\{\sum_{i=1}^{k}(p_i'-p_i)\,\bar{y}_{N_i}\right\}^2 = V\left\{\sum_{i=1}^{k} p_i'\bar{y}_{N_i}\right\}$$

$$= \frac{N-Q}{N}\,\frac{1}{Q}\,\frac{1}{N-1}\sum_{i=1}^{k} N_i\,(\bar{y}_{N_i}-\bar{y}_N)^2$$

$$= \frac{N-Q}{N-1}\,\frac{1}{Q}\sum_{i=1}^{k} p_i\,(\bar{y}_{N_i}-\bar{y}_N)^2$$

Now, (113) can be rewritten as

$$V\left(\sum_{i=1}^{k} p_i'\bar{y}_{n_i}\right) = \sum_{i=1}^{k}\left(\frac{1}{n_i}-\frac{1}{N_i}\right)p_i^2\mathrm{S}_i^2$$
$$+\frac{N-Q}{N-1}\,\frac{1}{Q}\left\{\sum_{i=1}^{k}\left(\frac{1}{n_i}-\frac{1}{N_i}\right)p_i\,(1-p_i)\,\mathrm{S}_i^2\right.$$
$$\left.+\sum_{i=1}^{k} p_i\,(\bar{y}_{N_i}-\bar{y}_N)^2\right\} \tag{114}$$

The first term on the right-hand side is clearly the variance of the mean of a stratified sample when the strata sizes are known. The effect of determining the strata sizes from a sample is thus to increase the variance of the estimate by

$$\frac{N-Q}{N-1}\,\frac{1}{Q}\left\{\sum_{i=1}^{k}\left(\frac{1}{n_i}-\frac{1}{N_i}\right)p_i\,(1-p_i)\,\mathrm{S}_i^2+\sum_{i=1}^{k}p_i\,(\bar{y}_{N_i}-\bar{y}_N)^2\right\} \tag{115}$$

When finite multipliers can be ignored, the effect is to increase the variance by approximately

$$\frac{\sigma_b^2}{Q} \tag{116}$$

where

$$\sigma_b^2 = \sum_{i=1}^{k} p_i\,(\bar{y}_{N_i}-\bar{y}_N)^2 \tag{117}$$

since the first term in (115) will be small relative to the second.

Compared to simple random sampling, on the other hand, the procedure will lead to gain in precision. Thus, for the case when n_i is proportional to $p_i\mathrm{S}_i$, and ignoring finite multipliers, (114) reduces to

$$V\left(\sum_{i=1}^{k} p_i'\bar{y}_{n_i}\right) = \left(1 - \frac{1}{Q}\right)\frac{1}{n}\left(\sum_{i=1}^{k} p_i\mathrm{S}_i\right)^2 + \frac{1}{Qn}\left(\sum_{i=1}^{k} \mathrm{S}_i\right)\left(\sum_{i=1}^{k} p_i\mathrm{S}_i\right) + \frac{\sigma_b{}^2}{Q}$$

$$\cong \frac{1}{n}\left(\sum_{i=1}^{k} p_i\mathrm{S}_i\right)^2 + \frac{1}{Q}\,\sigma_b{}^2 \qquad (118)$$

If the sample were chosen by the procedure of simple random sampling without stratification, the variance of the estimate would be

$$V(\bar{y}_n) = \frac{N-n}{N}\,\frac{\mathrm{S}^2}{n}$$

$$\cong \frac{\sum\limits_{i=1}^{k} p_i\mathrm{S}_i{}^2 + \sigma_b{}^2}{n} \qquad (119)$$

The reduction in variance is, therefore, given by

$$V_{US} - V_{ds} = \frac{1}{n}\sum_{i=1}^{k} p_i\,(\mathrm{S}_i - \bar{\mathrm{S}}_w)^2 + \left(\frac{1}{n} - \frac{1}{Q}\right)\sigma_b{}^2 \qquad (120)$$

where the letters '*ds*' stand for double sampling and, as before, $\bar{\mathrm{S}}_w = \sum\limits_{i=1}^{k} p_i\mathrm{S}_i$. If Q is large relative to n, the reduction in variance will approximate to the difference between the variances of an unstratified simple random sample and that of a stratified sample with Neyman allocation (*vide* Section 3*a*.5). In other words, when Q is large, the procedure of double sampling will be approximately equivalent to stratified sampling when strata sizes are accurately known.

This comparison of the precision of single with double sampling regardless of the cost of the survey is not, however, of practical value. What is of interest is to know whether the reduction in variance would be worth the extra expenditure on the preliminary sample. Alternatively, we can consider the problem as one of choosing, for a fixed cost, say C_0, between two procedures of sampling, namely, (i) a single sample drawn by the procedure of simple random sampling without stratification, and (ii) double sampling. The problem clearly envisages three steps, namely, (*a*) determining the optimum design for each of the two procedures of sampling, (*b*) obtaining the variances of the estimates for the optimum designs, and (*c*) comparison of the two variances.

We shall consider the simplest case in which the cost of each sample is proportional to its size so that the total cost of the survey, using the procedure of double sampling, is represented by

$$c_1 Q + c_2 \left(\sum_{i=1}^{k} n_i \right) \tag{121}$$

and that of a single sample of size, say n', drawn by procedure (i) by

$$c_2 n' \tag{122}$$

where c_1 is the cost per unit of the preliminary sample and c_2 the cost per unit for the main sample. Obviously c_1 will be smaller than c_2, for unless this were so, a preliminary sample would be ruled out altogether. For procedure (i) the optimum design is clearly the one for which the size of the single sample is given by

$$n' = \frac{C_0}{c_2}$$

The variance of the estimate based on a single sample is, therefore, given by

$$\frac{N - \dfrac{C_0}{c_2}}{N \dfrac{C_0}{c_2}} \mathbf{S}^2 \tag{123}$$

or, neglecting finite correction factors, by

$$\frac{\mathbf{S}^2 c_2}{C_0} \tag{124}$$

The variance of the estimate based on the procedure of double sampling given in (114) depends on Q and n_i $(i = 1, 2, \ldots, k)$. The optimum values of Q and n_i are those for which this variance is minimum for given C_0. Owing to the complexity of the algebra, we shall attempt only an approximate solution by ignoring all finite multipliers and assuming that

$$\frac{p_i(1-p_i)}{Q}$$

is negligible in comparison with p_i^2. Hence (114) reduces to

$$V\left(\sum_{i=1}^{k} p_i'\bar{y}_{n_i}\right) \cong \sum_{i=1}^{k} \frac{p_i^2S_i^2}{n_i} + \frac{\sigma_b^2}{Q} \tag{125}$$

This is to be minimized subject to the condition

$$C_0 = c_1Q + c_2 \sum_{i=1}^{k} n_i \tag{126}$$

Using the method given in the previous sections, we form the function ϕ given by

$$\phi = \sum_{i=1}^{k} \frac{p_i^2S_i^2}{n_i} + \frac{\sigma_b^2}{Q} + \mu\left(c_1Q + c_2\sum_{i=1}^{k} n_i\right) \tag{127}$$

and note that the right-hand side can be written as

$$\phi = \sum_{i=1}^{k} \left(\frac{p_i^2S_i^2}{n_i} + \mu c_2 n_i\right) + \frac{\sigma_b^2}{Q} + \mu c_1 Q$$

$$= \sum_{i=1}^{k} \left(\frac{p_iS_i}{\sqrt{n_i}} - \sqrt{\mu c_2 n_i}\right)^2 + \left(\frac{\sigma_b}{\sqrt{Q}} - \sqrt{\mu c_1 Q}\right)^2$$

$$+ 2\sqrt{\mu c_2} \sum_{i=1}^{k} p_iS_i + 2\sigma_b \sqrt{\mu c_1} \tag{128}$$

so that for optimum values

$$n_i = \frac{p_iS_i}{\sqrt{\mu c_2}} \qquad (i = 1, 2, \ldots, k) \tag{129}$$

and

$$Q = \frac{\sigma_b}{\sqrt{\mu c_1}} \tag{130}$$

The value of μ will depend upon whether the variance is minimized for fixed cost or the cost is minimized for fixed variance. In the former case, which is the one under consideration here, we obtain from (126), (129) and (130)

$$C_0 = \sqrt{\frac{c_1}{\mu}}\,\sigma_b + \sqrt{\frac{c_2}{\mu}}\sum_{i=1}^{k} p_i \mathbf{S}_i$$

giving us

$$\frac{1}{\sqrt{\mu}} = \frac{C_0}{\left(\sqrt{c_1}\,\sigma_b + \sqrt{c_2}\sum\limits_{i=1}^{k} p_i \mathbf{S}_i\right)} \tag{131}$$

Hence

$$n_i = \frac{p_i \mathbf{S}_i}{\sqrt{c_2}}\,\frac{C_0}{\left(\sqrt{c_1}\,\sigma_b + \sqrt{c_2}\sum\limits_{i=1}^{k} p_i \mathbf{S}_i\right)} \tag{132}$$

and

$$Q = \frac{\sigma_b}{\sqrt{c_1}}\,\frac{C_0}{\left(\sqrt{c_1}\,\sigma_b + \sqrt{c_2}\sum\limits_{i=1}^{k} p_i \mathbf{S}_i\right)} \tag{133}$$

Substituting for Q and n_i in the expression for the variance given in (125), we have

$$V\left(\sum_{i=1}^{k} p_i' \bar{y}_{n_i}\right) \cong \frac{\left(\sqrt{c_1}\,\sigma_b + \sqrt{c_2}\sum\limits_{i=1}^{k} p_i \mathbf{S}_i\right)^2}{C_0} \tag{134}$$

Now, for a fixed cost, double sampling would lead to higher precision if (134) is less than (124), that is, if

$$\frac{\left(\sqrt{c_1}\,\sigma_b + \sqrt{c_2}\sum\limits_{i=1}^{k} p_i \mathbf{S}_i\right)^2}{C_0} < \mathbf{S}^2\,\frac{c_2}{C_0}$$

i.e., if

$$\sqrt{c_1}\,\sigma_b + \sqrt{c_2}\sum_{i=1}^{k} p_i \mathbf{S}_i < \mathbf{S}\,\sqrt{c_2}$$

i.e., if

$$c_1 < c_2\,\frac{\left(\mathbf{S} - \sum\limits_{i=1}^{k} p_i \mathbf{S}_i\right)^2}{{\sigma_b}^2} \tag{135}$$

where $\left(\mathrm{S} - \sum_{i=1}^{k} p_i \mathrm{S}_i\right)^2 \Big/ \sigma_b^2$ will usually be a positive proper fraction for efficient systems of stratification.

Taking, for illustration, the case when

$$\mathrm{S}_i = \mathrm{S}_w = 1 \ (i = 1, 2, \ldots, k), \quad \sigma_b = 2$$

so that

$$\mathrm{S}^2 = 5$$

equation (135) reduces to

$$c_1 < \frac{1 \cdot 5}{4} c_2 = \frac{3}{8} c_2$$

In other words, if the cost per unit of surveying the preliminary sample is less than one-third the cost per unit of surveying the main sample, double sampling would be a more efficient procedure to adopt than simple random sampling without stratification.

Example 3.1

Table 3.1 presents the summary of data for complete census of all the 340 villages in Ghaziabad Subdivision. The villages were stratified by size of their agricultural area into four strata as shown in col. 2 of Table 3.1. The numbers of villages in the different strata are given in col. 3. The population values of the strata means for the area under wheat ($\bar{y}_{N_i}$) and those of the standard deviations for the area under wheat (S_{w_i}) and for the agricultural area (S_{a_i}) are given in the subsequent columns.

Calculate the sampling variance of the estimated area under wheat for a sample of 34 villages:

(1) if the villages are selected by the method of simple random sampling without stratification;

(2) if the villages are selected by the method of simple random sampling within each stratum, and allocated in proportion to (i) the sizes of the strata (N_i), (ii) the products $N_i\mathrm{S}_{w_i}$, and (iii) the products $N_i\mathrm{S}_{a_i}$.

TABLE 3.1

Strata Means and Standard Deviations of Areas for Villages in Ghaziabad Subdivision

Stratum Number	Size of Village in Bighas*	N_i	$\bar{y}_{N_i}$	S_{wi}	S_{ai}
(1)	(2)	(3)	(4)	(5)	(6)
1	0– 500	63	112·1	56·3	129·6
2	501–1500	199	276·7	116·4	267·0
3	1501–2500	53	558·1	186·0	276·1
4	>2500	25	960·1	361·3	982·2

* 1 Bigha = $\frac{5}{8}$ acre.

1. *Simple Random Sampling Without Stratification*

We have from (29)

$$S^2 = \frac{1}{N-1}\left[\sum_{i=1}^{k}(N_i - 1)\,S_{w_i}{}^2 + \sum_{i=1}^{k} N_i\,(\bar{y}_{N_i} - \bar{y}_N)^2\right]$$

$$= \frac{1}{N-1}\left[\sum_{i=1}^{k} N_i S_{w_i}{}^2 - \sum_{i=1}^{k} S_{w_i}{}^2 + \sum_{i=1}^{k} N_i \bar{y}_{N_i}{}^2 - \frac{\left(\sum\limits_{i=1}^{k} N_i \bar{y}_{N_i}\right)^2}{N}\right]$$

The relevant calculations are shown in Table 3.2. On substituting, we obtain

$$S^2 = \frac{1}{339}\,[7994000 - 182000 + 55577000 - 39372000]$$

$$= 70850$$

TABLE 3.2

Calculation of the Sampling Variance in Example 3.1

Stratum Number	N_i	S_{wi}	S_{wi}^2	$N_iS_{wi}^2$	$\bar{y}_{Ni}$	$N_i\bar{y}_{Ni}$	$N_i\bar{y}_{Ni}^2$	N_iS_{wi}	S_{ai}	N_iS_{ai}	$\frac{N_iS_{wi}^2}{S_{ai}}$
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
1	63	56·3	3170	200000	112·1	7060	791000	3550	129·6	8160	1540
2	199	116·4	13550	2696000	276·7	55060	15235000	23160	267·0	53130	10100
3	53	186·0	34600	1834000	558·1	29580	16509000	9860	276·1	14630	6640
4	25	361·3	130540	3264000	960·1	24000	23042000	9030	982·2	24560	3320
	340		181860	7994000		115700	55577000	45600		100480	21600

Hence

$$V(\bar{y}_n)_{US} = \frac{N-n}{N} \cdot \frac{1}{n} \mathbf{S}^2$$

$$= \frac{306}{340 \times 34} \times 70850$$

$$= 1875$$

2. (i) *Proportional Allocation*

We have

$$V(\bar{y}_w)_P = \frac{N-n}{N^2 n} \sum_{i=1}^{k} N_i \mathbf{S}_{wi}{}^2$$

$$= \frac{306}{(340)^2 \times 34} \times 7994000$$

$$= 622$$

2. (ii) *Neyman Allocation*

The allocation of the sample to the different strata will be in proportion to $N_i\mathbf{S}_{w_i}$ shown in col. 9 of Table 3.2. On substituting in (22), we get

$$V(\bar{y}_w)_N = \frac{1}{N^2 n}\left(\sum_{i=1}^{k} N_i \mathbf{S}_{wi}\right)^2 - \frac{1}{N^2} \sum_{i=1}^{k} N_i \mathbf{S}_{wi}{}^2$$

$$= \frac{1}{(340)^2}\left[\frac{(45600)^2}{34} - 7994000\right]$$

$$= \frac{1}{115600}\,[61158000 - 7994000]$$

$$= 460$$

2. (iii) *Allocation Proportional to* $N_i\mathbf{S}_{a_i}$

The allocation of the sample will be proportional to $N_i\mathbf{S}_{a_i}$ given in col. 11 of Table 3.2. On substituting

$$n_i = \frac{nN_i\mathbf{S}_{ai}}{\left(\sum_{j=1}^{k} N_j \mathbf{S}_{aj}\right)}$$

in the formula for the variance of the mean of a stratified sample given in (7), we get

$$V(\bar{y}_w)_S = \frac{1}{N^2 n}\left(\sum_{i=1}^{k} N_i \mathrm{S}_{ai}\right)\left(\sum_{i=1}^{k} N_i \frac{\mathrm{S}_{wi}^2}{\mathrm{S}_{ai}}\right) - \frac{1}{N^2}\sum_{i=1}^{k} N_i \mathrm{S}_{wi}^2$$

$$= \frac{1}{(340)^2}\left\{\frac{1}{34} \times (100480)\,(21600) - 7994000\right\}$$

$$= \frac{1}{115600} \times 55840000$$

$$= 483$$

Now the *relative efficiency* of any given procedure of sampling (B) compared to that of a standard procedure (A) for the same size of sample is defined as the ratio of the inverse of their variances. Thus

$$R.E. = \frac{\dfrac{1}{V_B}}{\dfrac{1}{V_A}} = \frac{V_A}{V_B}$$

The values of the variance obtained above for the different procedures of sampling together with those of their relative efficiencies as compared with unstratified simple random sampling, are given in Table 3.3. The values of the relative efficiencies compared to proportional sampling are also shown in the table.

TABLE 3.3

Relative Efficiencies of Different Methods of Stratified Sampling

Method of Sampling	Variance (Bighas)^2	*R.E.* compared to Unstratified Sampling	*R.E.* compared to Proportional Sampling
Unstratified Simple Random Sampling	1875	..	..
Stratified:			
(i) Proportional	622	301%	..
(ii) Neymar	460	408%	135%
(iii) $n_i \propto N_i \mathrm{S}_{ai}$	483	388%	129%

It will be seen that stratified sampling reduces the sampling variance to nearly one-third its value in unstratified simple random sampling. Further, Neyman allocation improves the precision as compared to proportional allocation. The allocation of the sample in accordance with the Neyman principle as applied to a correlated character is seen to be almost as effective in improving the precision as the Neyman method applied to the character under study.

Example 3.2

A yield survey on paddy was carried out in Kegalle District (Ceylon) in Maha 1951–52 (Koshal, 1953). Twenty-eight villages were selected, distributed in the various strata approximately in proportion to the acreage under paddy. Three plots of 1/80 acre were harvested in each village. The values of the means and the mean squares of the village means for the different strata are given in Table 3.4. Obtain the estimate of the district mean yield by combining the strata means in proportion to the number of villages in the strata. Calculate its variance and hence estimate the efficiency of stratification as compared to unstratified simple random sampling, treating the village means as the true means of the respective villages.

TABLE 3.4

Crop-Cutting Experiments on Paddy, Kegalle District (Ceylon), Maha 1951–52

Means and Mean Squares of Village Mean Yields per Plot

Stratum Number	N_i	n_i	$\bar{y}_{ni}$ (Oz./Plot)	s_i^2 (Oz./Plot)^2
1	189	5	369	4330·9
2	242	7	301	14812·4
3	146	3	368	17309·0
4	178	3	171	1658·5
5	287	10	305	3452·7

The relevant calculations are given in Table 3.5. From col. 5 we have

$$\text{Est. } \bar{y}_N = \bar{y}_w = 301 \cdot 6$$

To obtain the variance of $\bar{y}_w$, we substitute in (57) and obtain

$$\begin{aligned}\text{Est. } V(\bar{y}_w)_S &= \sum_{i=1}^{k} \frac{p_i^2 s_i^2}{n_i} - \sum_{i=1}^{k} \frac{p_i^2 s_i^2}{N_i}\\ &= 298 \cdot 23 - 7 \cdot 57\\ &= 290 \cdot 66\end{aligned}$$

From (72), taking $N/(N-1) \cong 1$, we obtain

$$\begin{aligned}\text{Est. } V(\bar{y}_n)_{US} &\cong \frac{N-n}{Nn} \left\{ \sum_{i=1}^{k} p_i s_i^2 + \sum_{i=1}^{k} p_i \bar{y}_{n_i}^2 - \left(\sum_{i=1}^{k} p_i \bar{y}_{n_i} \right)^2 \right.\\ &\qquad \left. - \sum_{i=1}^{k} \left(\frac{p_i s_i^2}{n_i} - \frac{p_i^2 s_i^2}{n_i} \right) \right\}\\ &= \left(\frac{1}{28} - \frac{1}{1042} \right) \left\{ 7885 + 95300 - 90960 - 1646 + 298 \right\}\\ &= (0 \cdot 034755)\,(10900)\\ &= 379\end{aligned}$$

Hence

$$\begin{aligned}\text{Efficiency of stratification} &= \frac{V(\bar{y}_n)_{US}}{V(\bar{y}_w)_S}\\ &= \frac{379}{291}\\ &= 1 \cdot 30 \text{ or } 130\%\end{aligned}$$

TABLE 3.5

Crop-Cutting Experiments on Paddy, Kegalle District, Maha 1951–52

Calculation of the District Mean Yield and its Variance

Stratum Number	N_i (1)	n_i (2)	$\bar{y}_{n_i}$ (3)	p_i (4)	$p_i\bar{y}_{n_i}$ (5) = (3)×(4)	$p_i\bar{y}_{n_i}^2$ (6) = (3)×(5)	s_i^2 (7)
1	189	5	369	·181382	66·9	24700	4330·9
2	242	7	301	·232246	69·9	21000	14812·4
3	146	3	368	·140115	51·6	19000	17309·0
4	178	3	171	·170825	29·2	5000	1658·5
5	287	10	305	·275432	84·0	25600	3452·7
	1042	28			301·6	95300	

Stratum Number	$p_is_i^2$ (8) = (4)×(7)	$p_i^2s_i^2$ (9)=(4)×(8)	$\frac{p_is_i^2}{n_i}$ (10)=(8)÷(2)	$\frac{p_i^2s_i^2}{n_i}$ (11)=(9)÷(2)	$\frac{p_i^2s_i^2}{N_i}$ (12)=(9)÷(1)
1	785·55	142·48	157·1	28·50	0·754
2	3440·12	798·95	491·4	114·14	3·301
3	2425·25	339·81	808·4	113·27	2·327
4	283·31	48·40	94·4	16·13	0·272
5	950·98	261·93	95·1	26·19	0·913
	7885·21		1646·4	298·23	7·567

B. VARYING PROBABILITIES OF SELECTION

3*b*.1 Estimate of the Population Mean and its Sampling Variance

Lastly we shall consider the theory of stratified sampling when sampling units within a stratum are selected with replacement with varying probabilities of selection. Let

$$P_{ij}\,(j=1, 2, \ldots, N_i;\ i=1, 2, \ldots, k)$$

denote the selection probability assigned to the *j*-th unit in the *i*-th stratum. Clearly, then, in virtue of the results in Section 2*b*.2, $\bar{z}_{n_i}$ defined by

$$\bar{z}_{n_i} = \frac{1}{n_i} \sum_{j}^{n_i} z_{ij}$$

$$= \frac{1}{n_i} \sum_{j}^{n_i} \frac{1}{N_i} \frac{y_{ij}}{P_{ij}} \tag{136}$$

will provide an unbiased estimate of the population mean for the i-th stratum, namely, $\bar{y}_{N_i}$; and its sampling variance will be given by

$$V\ (\bar{z}_{n_i}) = \frac{\sigma_{iz}{}^2}{n_i} \tag{137}$$

where

$$\sigma_{iz}{}^2 = \sum_{j=1}^{N_i} P_{ij}\,(z_{ij} - \bar{z}_{i.}\,)^2 \tag{138}$$

$$\bar{z}_{i.} = \sum_{j=1}^{N_i} P_{ij} z_{ij} = \bar{y}_{N_i} \tag{139}$$

and

$$\bar{z}_{..} = \sum_{i=1}^{k} p_i \bar{z}_{i.} = \bar{y}_N \tag{140}$$

An estimate of the population mean $\bar{y}_N$ will be the weighted mean $\bar{z}_w$ given by

$$\bar{z}_w = \frac{1}{N} \sum_{i=1}^{k} N_i \bar{z}_{n_i}$$

$$= \sum_{i=1}^{k} p_i \bar{z}_{n_i} \tag{141}$$

which is easily seen to provide an unbiased estimate of $\bar{y}_N$. For,

$$E\ (\bar{z}_w) = \sum_{i=1}^{k} p_i E\,(\bar{z}_{n_i})$$

$$= \sum_{i=1}^{k} p_i \frac{1}{n_i} \sum_{j}^{n_i} E\,(z_{ij})$$

$$= \sum_{i=1}^{k} p_i \bar{y}_{N_i}$$

$$= \bar{y}_N \tag{142}$$

To obtain the sampling variance of $\bar{z}_w$, we have

$$\begin{aligned} V(\bar{z}_w) &= E[\bar{z}_w - E(\bar{z}_w)]^2 \\ &= E\left(\sum_{i=1}^{k} p_i \bar{z}_{n_i} - \sum_{i=1}^{k} p_i \bar{z}_{i.}\right)^2 \\ &= E\left\{\sum_{i=1}^{k} p_i (\bar{z}_{n_i} - \bar{z}_{i.})\right\}^2 \\ &= E\left\{\sum_{i=1}^{k} p_i^2 (\bar{z}_{n_i} - \bar{z}_{i.})^2 + \sum_{i \neq i'=1}^{k} p_i p_{i'} (\bar{z}_{n_i} - \bar{z}_{i.}) \right. \\ &\qquad \left. \times (\bar{z}_{n_{i'}} - \bar{z}_{i'.})\right\} \\ &= \sum_{i=1}^{k} p_i^2 E(\bar{z}_{n_i} - \bar{z}_{i.})^2 + \sum_{i \neq i'=1}^{k} p_i p_{i'} E(\bar{z}_{n_i} - \bar{z}_{i.}) \\ &\qquad \times E(\bar{z}_{n_{i'}} - \bar{z}_{i'.}) \end{aligned}$$

since samples are drawn independently from the i-th and the i'-th strata. Hence,

$$V(\bar{z}_w) = \sum_{i=1}^{k} p_i^2 \frac{\sigma_{iz}^2}{n_i} \tag{143}$$

in virtue of (137).

Using Section 2b.3, an estimate of $V(\bar{z}_w)$ will be provided by

$$\text{Est. } V(\bar{z}_w) = \sum_{i=1}^{k} p_i^2 \frac{s_{iz}^2}{n_i} \tag{144}$$

where s_{iz}^2 denotes the mean square between z's in the sample for the i-th stratum, and is defined by

$$s_{iz}^2 = \frac{1}{n_i - 1} \sum_{j}^{n_i} (z_{ij} - \bar{z}_{n_i})^2$$

3b.2 Allocation of Sample among Different Strata

The variance of the estimate, apart from the population constants p_i and σ_{iz}, is seen to depend upon the allocation of the sample among the different strata. The cost of a survey will likewise depend upon the values of n_i. The principle of determining the optimum values of n_i, as stated in Section 3a.3, is to maximize the precision for given cost or minimize the cost for given precision. We shall illustrate the principle for the simple case for which the cost of the survey is represented by

$$C = \sum_{i=1}^{k} c_i n_i \tag{145}$$

where c_i is the cost of collecting the information per unit in the i-th stratum.

Let

$$\begin{aligned} \phi &= V(\bar{z}_w) + \mu C \\ &= \sum_{i=1}^{k} p_i^2 \frac{\sigma_{iz}^2}{n_i} + \mu \left(\sum_{i=1}^{k} c_i n_i \right) \end{aligned}$$

where μ is a constant.

Clearly $V(\bar{z}_w)$ is minimum for fixed cost, say C_0, or C is minimum for fixed variance, say V_0, when ϕ is a minimum. Now ϕ can be written as

$$\phi = \sum_{i=1}^{k} \left\{ \frac{p_i \sigma_{iz}}{\sqrt{n_i}} - \sqrt{\mu c_i n_i} \right\}^2 + 2\sqrt{\mu} \sum_{i=1}^{k} p_i \sigma_{iz} \sqrt{c_i} \tag{146}$$

It follows that ϕ is minimum when each of the square terms on the right-hand side of (146) is zero, or in other words

$$n_i = \frac{1}{\sqrt{\mu}} \cdot \frac{p_i \sigma_{iz}}{\sqrt{c_i}} \qquad (i = 1, 2, \ldots, k) \tag{147}$$

The constant of proportionality $1/\sqrt{\mu}$ is determined so as to satisfy the condition of fixed cost or fixed variance. In the former case, we substitute for n_i from (147) in (145) and obtain

$$\frac{1}{\sqrt{\mu}} = \frac{C_0}{\sum\limits_{i=1}^{k} p_i \sigma_{iz} \sqrt{c_i}} \tag{148}$$

Hence, from (147), we get

$$n_i = \frac{p_i \sigma_{iz} C_0}{\sqrt{c_i} \left(\sum\limits_{i=1}^{k} p_i \sigma_{iz} \sqrt{c_i} \right)} \tag{149}$$

which is seen to be identical in form with (13) in Section 3*a*.3.

When $c_i = c$, or, in other words, when the total size of sample is fixed, the optimum allocation is given by

$$n_i = n \frac{p_i \sigma_{iz}}{\sum\limits_{i=1}^{k} p_i \sigma_{iz}} \tag{150}$$

For the alternative approach in which the cost is minimized for given precision, the reader may verify that the value of n_i is given by

$$n_i = \frac{p_i \sigma_{iz}}{\sqrt{c_i}} \cdot \frac{\sum\limits_{i=1}^{k} p_i \sigma_{iz} \sqrt{c_i}}{V_0} \tag{151}$$

where V_0 stands for the value of the variance with which it is desired to estimate the mean. Comparing (151) with (17), we notice that the optimum allocation is governed by the same considerations as those mentioned in Section 3*a*.3 for simple random sampling.

3*b*.3 Variance of the Estimate under (i) Optimum Allocation, and (ii) Proportional Allocation when the Total Size of Sample is Fixed

For n fixed, the optimum allocation is given by (150). Substituting for n_i from (150) in (143), we get

$$V(\bar{z}_w)_N = \frac{1}{n} \left\{ \sum_{i=1}^{k} p_i \sigma_{iz} \right\}^2 \tag{152}$$

For proportional allocation we substitute $n_i = np_i$ in (143) and obtain

$$V(\bar{z}_w)_P = \frac{1}{n} \sum_{i=1}^{k} p_i {\sigma_{iz}}^2 \tag{153}$$

Now (153) can be expressed as

$$V(\bar{z}_w)_P = \frac{1}{n} \left[\left\{ \sum_{i=1}^{k} p_i \sigma_{iz} \right\}^2 + \sum_{i=1}^{k} p_i (\sigma_{iz} - \bar{\sigma}_{wz})^2 \right] \tag{154}$$

where

$$\bar{\sigma}_{wz} = \sum_{i=1}^{k} p_i \sigma_{iz} \tag{155}$$

It follows that the efficiency of optimum over proportional allocation is due wholly to the variation among the strata standard deviations. If the σ_{iz} are all equal, the two systems of allocation become equally efficient.

3*b*.4 Comparison of Stratified with Unstratified Sampling

If a sample of n is selected as an unstratified sample with replacement with selection probabilities P_l $(l = 1, 2, \ldots, N)$, then we have seen that

$$\bar{z}_n = \frac{1}{n} \sum^{n} z_l$$

$$= \frac{1}{n} \sum^{n} \frac{y_l}{NP_l} \tag{156}$$

provides an unbiased estimate of the population mean $\bar{y}_N$, and that its sampling variance is given by

$$V(\bar{z}_n) = \frac{\sigma_z^2}{n} \tag{157}$$

where

$$\sigma_z^2 = \sum_{l=1}^{N} P_l (z_l - \bar{z}_{..})^2 \tag{158}$$

and

$$\bar{z}_{..} = \sum_{l=1}^{N} P_l z_l = \bar{y}_N \tag{159}$$

If the sample of n is selected as a stratified sample with n_i units from the i-th stratum with selection probabilities proportional to P_l given by

$$P_{ij} = \frac{P_l}{\sum^{N_i} P_l} \qquad (j = 1, 2, \ldots, N_i) \tag{160}$$

where $\sum^{N_i}$ denotes summation over the N_i units in the i-th stratum, then

$$\bar{z}_w = \sum_{i=1}^{k} p_i \bar{z}_{n_i} \tag{161}$$

provides an unbiased estimate of $\bar{y}_N$, with its sampling variance given by

$$V(\bar{z}_w) = \sum_{i=1}^{k} p_i^2 \frac{\sigma_{iz}^2}{n_i} \tag{162}$$

where

$$\sigma_{iz}{}^2 = \overset{N_i}{\underset{j=1}{\Sigma}} P_{ij}\,(z_{ij} - \bar{z}_{i.})^2 \tag{163}$$

For purposes of comparing (157) and (162), we note that

$$\begin{aligned} P_l &= P_{ij} \overset{N_i}{\Sigma}{}' P_l \\ &= P_{ij}P_{i.} \end{aligned} \tag{164}$$

where

$$P_{i.} = \overset{N_i}{\Sigma}{}' P_l \tag{165}$$

Also

$$\begin{aligned} z_l &= \frac{y_l}{NP_l} \\ &= \frac{y_l}{NP_{ij}P_{i.}} \\ &= \frac{p_i}{P_{i.}}\, z_{ij} \end{aligned} \tag{166}$$

We may, therefore, write

$$\begin{aligned} \sigma_z{}^2 &= \sum_{l=1}^{N} P_l\,(z_l - \bar{z}_{..})^2 \\ &= \sum_{i=1}^{k} P_{i.} \sum_{j=1}^{N_i} P_{ij} \left(\frac{p_i}{P_{i.}}\ z_{ij} - \bar{z}_{..}\right)^2 \\ &= \sum_{i=1}^{k} P_{i.} \sum_{j=1}^{N_i} P_{ij} \left(\frac{p_i}{P_{i.}}\ z_{ij}\ -\ \frac{p_i}{P_{i.}}\ \bar{z}_{i.}\ +\ \frac{p_i}{P_{i.}}\ \bar{z}_{i.} - \bar{z}_{..}\right)^2 \\ &= \sum_{i=1}^{k} P_{i.} \sum_{j=1}^{N_i} P_{ij} \left\{\left(\frac{p_i}{P_{i.}}\ z_{ij}\ -\ \frac{p_i}{P_{i.}}\ \bar{z}_{i.}\right)^2 + \left(\frac{p_i}{P_{i.}}\ \bar{z}_{i.} - \bar{z}_{..}\right)^2\right\} \\ &= \sum_{i=1}^{k} \frac{p_i^2}{P_{i.}} \sum_{j=1}^{N_i} P_{ij}\,(z_{ij} - \bar{z}_{i.})^2 + \sum_{i=1}^{k} P_{i.} \left(\frac{p_i}{P_{i.}}\bar{z}_{i.} - \bar{z}_{..}\right)^2 \\ &= \sum_{i=1}^{k} \frac{p_i{}^2}{P_{i.}}\ \sigma_{iz}{}^2 + \sum_{i=1}^{k} P_{i.} \left(\frac{p_i}{P_{i.}}\ \bar{z}_{i.} - \bar{z}_{..}\right)^2 \end{aligned} \tag{167}$$

Hence, from (157) and (162), we get

$$V_{US} - V_S = \sum_{i=1}^{k} p_i^2 \sigma_{iz}^2 \left(\frac{1}{nP_{i.}} - \frac{1}{n_i}\right) + \frac{1}{n}\sum_{i=1}^{k} P_{i.}\left(\frac{p_i}{P_{i.}}\bar{z}_{i.} - \bar{z}_{..}\right)^2 \tag{168}$$

Now the first term in (168) can be positive, zero or negative. It is zero when the sample is so allocated among the different strata that

either

$$n_i \propto P_{i.}$$

$$= n \sum^{N_i} P_i \tag{169}$$

or

$$n_i \propto \frac{p_i^2 \sigma_{iz}^2}{P_{i.}}$$

$$= n \frac{p_i^2 \dfrac{\sigma_{iz}^2}{P_{i.}}}{\sum_{i=1}^{k} p_i^2 \dfrac{\sigma_{iz}^2}{P_{i.}}} \tag{170}$$

The variance of an unstratified sample in this case is reduced by

$$V_{US} - V_P = \frac{1}{n} \sum_{i=1}^{k} P_{i.} \left(\frac{p_i}{P_{i.}} \bar{z}_{i.} - \bar{z}_{..}\right)^2 \tag{171}$$

The first term in (168) is positive when the sample is allocated according to the Neyman principle of allocation in (150). We have for this case

$$V_{US} - V_N = \frac{1}{n}\left\{\sum_{i=1}^{k} P_{i.}\left(\frac{p_i \sigma_{iz}}{P_{i.}} - \bar{\sigma}_{wz}\right)^2\right\}$$
$$+ \frac{1}{n}\left\{\sum_{i=1}^{k} P_{i.}\left(\frac{p_i}{P_{i.}} \bar{z}_{i.} - \bar{z}_{..}\right)^2\right\} \tag{172}$$

The efficiency of a stratified sample will decrease as the allocation will depart from the Neyman principle and a point may be reached where the first term in (168) will not only be negative but larger in magnitude than the second term, thus making an unstratified sample more efficient than a stratified sample.

For the special case when $P_{i.} = p_i$, (168) takes the form

$$\{V_{US} - V_S\}_{(P_{i.}=p_i)} = \sum_{i=1}^{k} p_i^2 \sigma_{iz}^2 \left(\frac{1}{np_i} - \frac{1}{n_i}\right) + \frac{1}{n} \sum_{i=1}^{k} p_i (\bar{z}_{i.} - \bar{z}_{..})^2 \qquad (173)$$

and in addition when the allocation is optimum, in accordance with the Neyman principle, the reduction in variance is given by

$$\{V_{US} - V_N\}_{(P_{i.}=p_i)} = \frac{1}{n} \left\{\sum_{i=1}^{k} p_i (\sigma_{iz} - \bar{\sigma}_{wz})^2\right\} + \frac{1}{n} \left\{\sum_{i=1}^{k} p_i (\bar{z}_{i.} - \bar{z}_{..})^2\right\} \qquad (174)$$

3*b*.5 Evaluation of the Change in Variance due to Stratification

In this section we shall evaluate, from a given stratified sample, the difference between the variance of an unstratified and a stratified sample. We write from (168)

$$\text{Est. } \{V_{US} - V_S\} = \sum_{i=1}^{k} p_i^2 \hat{\sigma}_{iz}^2 \left(\frac{1}{nP_{i.}} - \frac{1}{n_i}\right) + \frac{1}{n} \text{Est.} \left\{\sum_{i=1}^{k} \frac{p_i^2 \bar{z}_{i.}^2}{P_{i.}} - \bar{z}_{..}^2\right\} \qquad (175)$$

Now, from (137),

$$V(\bar{z}_{n_i}) = E(\bar{z}_{n_i}^2) - \bar{z}_{i.}^2 = \frac{\sigma_{iz}^2}{n_i}$$

whence

$$\text{Est. } \sum_{i=1}^{k} \frac{p_i^2 \bar{z}_{i.}^2}{P_{i.}} = \sum_{i=1}^{k} \frac{p_i^2 \bar{z}_{n_i}^2}{P_{i.}} - \sum_{i=1}^{k} \frac{p_i^2 \hat{\sigma}_{iz}^2}{P_{i.} n_i} \qquad (176)$$

Also, from (143),

$$V(\bar{z}_w) = E(\bar{z}_w^2) - \bar{z}_{..}^2 = \sum_{i=1}^{k} p_i^2 \frac{\sigma_{iz}^2}{n_i}$$

whence

$$\text{Est. } \bar{z}_{..}^2 = \bar{z}_w^2 - \sum_{i=1}^{k} p_i^2 \frac{\hat{\sigma}_{iz}^2}{n_i} \tag{177}$$

Subtracting (177) from (176), we get

$$\text{Est. } \left\{\sum_{i=1}^{k} \frac{p_i^2 \bar{z}_{i.}^2}{P_{i.}} - \bar{z}_{..}^2\right\} = \sum_{i=1}^{k} \frac{p_i^2 \bar{z}_{n_i}^2}{P_{i.}} - \bar{z}_w^2$$

$$- \sum_{i=1}^{k} p_i^2 \frac{\hat{\sigma}_{iz}^2}{n_i} \left(\frac{1}{P_{i.}} - 1\right) \tag{178}$$

On substituting from (178) in (175), we, therefore, obtain

$$\text{Est. } \{V_{US} - V_S\} = \sum_{i=1}^{k} p_i^2 \hat{\sigma}_{iz}^2 \left(\frac{1}{nP_{i.}} - \frac{1}{n_i}\right)$$

$$+ \frac{1}{n} \left\{\sum_{i=1}^{k} \frac{p_i^2 \bar{z}_{n_i}^2}{P_{i.}} - \bar{z}_w^2\right\} - \frac{1}{n} \sum_{i=1}^{k} \frac{p_i^2 \hat{\sigma}_{iz}^2}{n_i} \left(\frac{1}{P_{i.}} - 1\right) \tag{179}$$

When $P_{i.} = p_i$, we get

$$\text{Est. } \{V_{US} - V_S\}_{(P_{i.} = p_i)} = \frac{1}{n} \sum_{i=1}^{k} p_i \hat{\sigma}_{iz}^2 - \sum_{i=1}^{k} p_i^2 \frac{\hat{\sigma}_{iz}^2}{n_i}$$

$$+ \frac{1}{n} \left\{\sum_{i=1}^{k} p_i (\bar{z}_{n_i} - \bar{z}_w)^2\right\} - \frac{1}{n} \sum_{i=1}^{k} p_i (1 - p_i) \frac{\hat{\sigma}_{iz}^2}{n_i} \tag{180}$$

which is seen to be identical in form with (73) after making N large in the latter.

REFERENCES

1. Neyman, J. (1934) .. "On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection," *Jour. Roy. Statist. Soc.*, **97**, 558–606.
2. Tschuprow, A. A. (1923) "On the Mathematical Expectation of the Moments of Frequency Distributions in the Case of Correlated Observations," *Metron*, **2**, No. 4.
3. Sukhatme, P. V. (1935) .. "Contribution to the Theory of the Representative Method," *Jour. Roy. Statist. Soc. Suppl.*, **2**, 253–68.
4. Evans, W. D. (1951) .. "On Stratification and Optimum Allocations," *Jour. Amer. Statist. Assoc.*, **46**, 95–104.
5. Neyman, J. (1938) .. "Contribution to the Theory of Sampling Human Populations," *Jour. Amer. Statist. Assoc.*, **33**, 101–16.
6. Koshal, R. S. (1953) .. "Report to the Government of Ceylon on Sample Survey of Rice," *E.T.A.P.*, Food and Agriculture Organisation of the United Nations, Rome.

CHAPTER IV

RATIO METHOD OF ESTIMATION

A. SAMPLING WITH EQUAL PROBABILITY OF SELECTION

4a.1 Introduction

In developing the theory of simple random sampling in the preceding chapters, we have considered only estimates based on simple arithmetic means of the observed values in the sample. In this and the next chapter, we shall consider other methods of estimation which make use of the ancillary information and which, under certain conditions, give more reliable estimates of the population values than those based on the simple averages. Two of these methods are of particular importance. They are: (i) the ratio method of estimation, and (ii) the regression method of estimation. In this chapter we shall consider the former.

4a.2 Notation and Definition of the Ratio Estimate

Let

y_i denote the value of the variate under study for the i-th unit of the population,

x_i the value of the ancillary variate on the same unit,

Y the total value of the y variate in the population,

X the total value of the x variate in the population,

$r_i = \dfrac{y_i}{x_i}$ the ratio of y to x for the i-th unit,

$\bar{r}_N = \frac{1}{N} \sum_{i=1}^{N} r_i$ the simple arithmetic mean of the ratios for all units in the population,

$\bar{r}_n = \frac{1}{n} \sum^{n} r_i$ the simple arithmetic mean of the ratios for the units in the sample,

$R_N = \frac{\bar{y}_N}{\bar{x}_N} = \frac{Y}{X}$ the ratio of the population mean of y to the population mean of x,

and

$R_n = \frac{\bar{y}_n}{\bar{x}_n} = \frac{\sum^{n} y_i}{\sum^{n} x_i}$ the corresponding ratio for the sample. (1)

R_n is said to provide an estimate of the population ratio R_N, and the product of R_n with X, given by

$$Y_R = R_n \cdot X \tag{2}$$

provides an estimate of the total value Y in the population. The estimate is known as the *ratio estimate* of the population total and its use presupposes the knowledge of X, the population total of x.

To take an example, y may denote the number of bullocks on a holding and x its area, the ratio R_n giving an estimate of the number of bullocks per acre of a holding in the population. The product of R_n with the total acreage of the holdings gives an estimate of the bullock population on the holdings. Or, again, y and x may denote the values of the character under study in two successive periods, *e.g.*, the acreage under a crop during the current and the census years. It will be shown in a subsequent section that by taking advantage of the correlation between y and x, the ratio method, under certain conditions, provides a more reliable estimate of the population value than the comparable estimate based on the simple arithmetic mean.

4*a*.3 Expected Value of the Ratio Estimate

At the outset it will be noticed that, unlike the estimate based on the simple arithmetic mean, in a ratio estimate the numerator and the denominator are both random variables. The derivation of the expected value of R_n, therefore, presents difficulties.

Let

$$y_i = \bar{y}_N + \epsilon_i$$

so that

$$\bar{y}_n = \bar{y}_N + \bar{\epsilon}_n \tag{3}$$

where

$$E(\bar{\epsilon}_n) = 0 \quad \text{and} \quad E(\bar{\epsilon}_n^2) = \frac{N-n}{N} \frac{\mathbf{S}_y^2}{n} \tag{4}$$

Similarly, let

$$x_i = \bar{x}_N + \epsilon_i'$$

so that

$$\bar{x}_n = \bar{x}_N + \bar{\epsilon}_n' \tag{5}$$

where

$$E(\bar{\epsilon}_n') = 0 \quad \text{and} \quad E(\bar{\epsilon}_n'^2) = \frac{N-n}{N} \frac{\mathbf{S}_x^2}{n} \tag{6}$$

To obtain the expected value of R_n, it is convenient to express it in terms of ϵ and ϵ'. We have, taking expectations,

$$E(R_n) = E\left\{\frac{\bar{y}_N\left(1 + \frac{\bar{\epsilon}_n}{\bar{y}_N}\right)}{\bar{x}_N\left(1 + \frac{\bar{\epsilon}_n'}{\bar{x}_N}\right)}\right\} \tag{7}$$

We shall now suppose that the x's are positive and n is sufficiently large, so that

$$\left|\frac{\bar{\epsilon}_n'}{\bar{x}_N}\right| < 1$$

Expanding* then

$$\left(1 + \frac{\bar{\epsilon}_n'}{\bar{x}_N}\right)^{-1}$$

as a series in $\bar{\epsilon}_n'$, we have

$$E(R_n) = R_N \cdot E\left\{1 + \frac{\bar{\epsilon}_n}{\bar{y}_N} - \frac{\bar{\epsilon}_n'}{\bar{x}_N} + \frac{\bar{\epsilon}_n'^2}{\bar{x}_N^2} - \frac{\bar{\epsilon}_n\bar{\epsilon}_n'}{\bar{y}_N\bar{x}_N}\right.$$
$$\left. + \frac{\bar{\epsilon}_n}{\bar{y}_N}\,\frac{\bar{\epsilon}_n'^2}{\bar{x}_N^2} - \frac{\bar{\epsilon}_n'^3}{\bar{x}_N^3} + \frac{\bar{\epsilon}_n'^4}{\bar{x}_N^4} - \frac{\bar{\epsilon}_n}{\bar{y}_N}\,\frac{\bar{\epsilon}_n'^3}{\bar{x}_N^3} + \ldots\right\} \tag{8}$$

Further, we shall assume that the contribution of terms involving powers in $\bar{\epsilon}_n$ and $\bar{\epsilon}_n'$ higher than the second to the value of $E(R_n)$ is negligible, being of the order of $1/n^\nu$ where $\nu > 1$. Denoting to a first approximation the expected value of R_n by $E_1(R_n)$, we may write

$$E_1(R_n) = R_N E\left\{1 + \frac{\bar{\epsilon}_n}{\bar{y}_N} - \frac{\bar{\epsilon}_n'}{\bar{x}_N} + \frac{\bar{\epsilon}_n'^2}{\bar{x}_N^2} - \frac{\bar{\epsilon}_n\bar{\epsilon}_n'}{\bar{y}_N\bar{x}_N}\right\} \tag{9}$$

* In a paper read before the meeting of the International Statistical Institute, 1951, J. C. Koop justified the expansion by using an ingenious device. He wrote

$$\sum^{n} y_i = \sum^{N} y_i - \sum^{N-n} y_i$$

so that

$$n\bar{y}_n = N\bar{y}_N - (N - n)\,\bar{y}_{N-n}$$

Similarly

$$n\bar{x}_n = N\bar{x}_N - (N - n)\,\bar{x}_{N-n}$$

Hence

$$R_n = R_N\left(1 - \frac{N-n}{N}\,\frac{\bar{y}_{N-n}}{\bar{y}_N}\right)\left(1 - \frac{N-n}{N}\,\frac{\bar{x}_{N-n}}{\bar{x}_N}\right)^{-1}$$

Clearly, when the x's are positive

$$\left|\frac{N-n}{N}\,\frac{\bar{x}_{N-n}}{\bar{x}_N}\right| < 1$$

Expanding

$$\left(1 - \frac{N-n}{N}\,\frac{\bar{x}_{N\;n}}{\bar{x}_N}\right)^{-1}$$

by Taylor's theorem and working out expectation term by term, he reached the same expression for the expected value of R_n as given in this and the next section.

Now

$$E(\bar{\epsilon}_n\bar{\epsilon}_n') = \frac{1}{n^2}\, E\left\{\left(\sum^{n}\epsilon_i\right)\left(\sum^{n}\epsilon_i'\right)\right\}$$

$$= \frac{1}{n^2}\, E\left\{\sum^{n}\epsilon_i\epsilon_i' + \sum_{i\neq j}^{n}\epsilon_i\epsilon_j'\right\}$$

$$= \frac{1}{n^2}\left\{\frac{n}{N}\sum_{i=1}^{N}\epsilon_i\epsilon_i' + \frac{n(n-1)}{N(N-1)}\sum_{i\neq j=1}^{N}\epsilon_i\epsilon_j'\right\}$$

in virtue of (47) of Section 2*a*.9,

$$= \frac{1}{nN}\sum_{i=1}^{N}\epsilon_i\epsilon_i' + \frac{(n-1)}{nN(N-1)}\left\{\left(\sum_{i=1}^{N}\epsilon_i\right)\left(\sum_{i=1}^{N}\epsilon_i'\right) - \sum_{i=1}^{N}\epsilon_i\epsilon_i'\right\}$$

$$= \frac{N-n}{Nn}\,\frac{1}{N-1}\left\{\sum_{i=1}^{N}\epsilon_i\epsilon_i'\right\}$$

$$= \frac{N-n}{N}\,\frac{1}{n}\,\rho\mathbf{S}_y\mathbf{S}_x \tag{10}$$

where ρ is the coefficient of correlation between y and x, given by

$$\rho = \frac{E(y_i-\bar{y}_N)(x_i-\bar{x}_N)}{\sqrt{E(y_i-\bar{y}_N)^2\cdot E(x_i-\bar{x}_N)^2}} \tag{11}$$

Substituting from (4), (6) and (10) in (9), we get

$$E_1(R_n) = R_N\left\{1 + \frac{N-n}{Nn}\left(\frac{\mathbf{S}_x^2}{\bar{x}_N^2} - \frac{\rho\mathbf{S}_y\mathbf{S}_x}{\bar{y}_N\bar{x}_N}\right)\right\}$$

$$= R_N\left\{1 + \frac{N-n}{Nn}\left(C_x^2 - \rho C_yC_x\right)\right\} \tag{12}$$

where

$$C_x = \frac{\mathbf{S}_x}{\bar{x}_N}$$

and

$$C_y = \frac{\mathbf{S}_y}{\bar{y}_N}$$

We notice that the expected value of R_n is not the population value R_N, showing that R_n is a biased estimate of R_N. Denoting to a first approximation the relative bias in a ratio estimate by B_1, we have

$$B_1 = \frac{E_1(R_n) - R_N}{R_N} = \frac{N-n}{N}\,\frac{1}{n}\,(C_x^2 - \rho C_y C_x) \tag{13}$$

When $C_y = C_x = C$, the expression for B_1 simplifies and we get for large N,

$$B_1 = \frac{C^2}{n}\,(1-\rho)$$

Thus for $C^2 = 0{\cdot}8$, $\rho = 0{\cdot}6$ and $n = 10$, the bias is seen to be a little over three per cent. of the population value of the ratio. The bias decreases as n increases, showing that the ratio estimate is a consistent estimate.* For n large and ρ sufficiently high, the bias will usually be negligible. The bias vanishes altogether when

$$C_x^2 = \rho C_y C_x$$

i.e.,

$$\bar{y}_N = \rho\,\frac{\mathbf{S}_y}{\mathbf{S}_x}\,\bar{x}_N$$

In other words, the bias vanishes when the regression of y on x is a straight line through the origin. This result is, in fact, general. For, let the regression of y on x be represented by, say,

$$E(y \mid x) = \beta x \tag{14}$$

Summing both sides of (14) over all units in the population, we obtain

$$\bar{y}_N = \beta \bar{x}_N$$

* The expression 'consistent estimate' is generally applied to those estimates from infinite populations which converge in probability to the population value.

or, in other words,

$$\beta = R_N \tag{15}$$

It follows that

$$E(R_n) = EE\left(\frac{\bar{y}_n}{\bar{x}_n}\,\middle|\,\begin{matrix}x_1\\ \vdots\\ x_n\end{matrix}\right)$$

$$= E\left\{\frac{\beta \bar{x}_n}{\bar{x}_n}\right\}$$

$$= \beta$$

$$= R_N$$

An important point concerning the magnitude of the bias which will be clear later on is that, in large samples, the bias in a ratio estimate is negligible as compared with the standard error of the estimate.

4*a*.4* Second Approximation to the Expected Value of the Ratio Estimate

In deriving the expected value of R_n in the previous section, we assumed that the contribution of terms involving powers in $\bar{\epsilon}_n$ and $\bar{\epsilon}_n'$ higher than the second is negligible. We shall now retain the terms in $\bar{\epsilon}_n$ and $\bar{\epsilon}_n'$ up to and including degree four, and proceed to obtain a better approximation to the expected value of R_n.

Taking expectations term by term in (8), and using $E_2(R_n)$ to denote the second approximation to the expected value of R_n, we write

$$E_2(R_n) = R_N\left[1 + \frac{E(\bar{\epsilon}_n'^2)}{\mu_{01}^2} - \frac{E(\bar{\epsilon}_n\bar{\epsilon}_n')}{\mu_{10}\mu_{01}} + \frac{E(\bar{\epsilon}_n\bar{\epsilon}_n'^2)}{\mu_{10}\mu_{01}^2} - \frac{E(\bar{\epsilon}_n'^3)}{\mu_{01}^3} + \frac{E(\bar{\epsilon}_n'^4)}{\mu_{01}^4} - \frac{E(\bar{\epsilon}_n\bar{\epsilon}_n'^3)}{\mu_{10}\mu_{01}^3}\right] \tag{16}$$

where $\mu_{10} = \bar{y}_N$ and $\mu_{01} = \bar{x}_N$. The evaluation of the terms on the right-hand side involves heavy algebra which is best dealt with by the method of bi-partitional functions. The relevant formulæ have been tabulated by the author (1944) and reproduced in the

Appendix to this chapter. Using then (6), (10) and the formulæ in the Appendix, and writing in terms of the moment notation given by

$$\mu_{a,\alpha} = \frac{1}{N}\sum_{i=1}^{N} \epsilon_i^{a}\epsilon_i'^{\alpha}$$

we have

$$\begin{aligned} E_2(R_n) = R_N \Bigg[1 &+ \frac{N-n}{N-1}\,\frac{1}{n}\left(\frac{\mu_{02}}{\mu_{01}^{2}} - \frac{\mu_{11}}{\mu_{10}\mu_{01}}\right) \\ &+ \frac{(N-n)(N-2n)}{(N-1)(N-2)}\,\frac{1}{n^2}\left(\frac{\mu_{12}}{\mu_{10}\mu_{01}^{2}} - \frac{\mu_{03}}{\mu_{01}^{3}}\right) \\ &+ \frac{(N-n)(N^2-6Nn+N+6n^2)}{(N-1)(N-2)(N-3)}\,\frac{1}{n^3} \\ &\qquad \times \left(\frac{\mu_{04}}{\mu_{01}^{4}} - \frac{\mu_{13}}{\mu_{10}\mu_{01}^{3}}\right) \\ &+ \frac{N(N-n)(N-n-1)}{(N-1)(N-2)(N-3)}\,\frac{3(n-1)}{n^3} \\ &\qquad \times \left(\frac{\mu_{02}^{2}}{\mu_{01}^{4}} - \frac{\mu_{11}\mu_{02}}{\mu_{10}\mu_{01}^{3}}\right)\Bigg] \end{aligned} \qquad (17)$$

It is seen that the contribution of higher order terms depends, apart from n, on the values of the moments and product-moments of the two variates. To get an idea of its magnitude, we shall suppose that N is large and that, further, the population follows a bivariate normal distribution, so that

$$\mu_{12} = 0 = \mu_{21};\ \mu_{03} = 0 = \mu_{30};\ \mu_{04} = 3\mu_{02}^{2};\ \mu_{13} = 3\mu_{11}\mu_{02}$$

We then have

$$\begin{aligned} E_2(R_n) &= R_N\left[1 + \frac{1}{n}\left(\frac{\mu_{02}}{\mu_{01}^{2}} - \frac{\mu_{11}}{\mu_{10}\mu_{01}}\right) + \frac{3}{n^2}\,\frac{\mu_{02}}{\mu_{01}^{2}} \right. \\ &\qquad\qquad \left. \times \left(\frac{\mu_{02}}{\mu_{01}^{2}} - \frac{\mu_{11}}{\mu_{10}\mu_{01}}\right)\right] \\ &= R_N\left[1 + \frac{1}{n}(C_x^2 - \rho C_y C_x) + \frac{3}{n^2}C_x^2(C_x^2 - \rho C_y C_x)\right] \\ &= R_N\left[1 + \frac{1}{n}(C_x^2 - \rho C_y C_x)\left(1 + \frac{3}{n}C_x^2\right)\right] \end{aligned} \qquad (18)$$

To a second approximation, the relative bias in a ratio estimate in samples of n from a large population can, therefore, be expressed as

$$B_2 = B_1\left(1 + \frac{3}{n}C_x^2\right) \tag{19}$$

Equation (19) shows that the contribution of the third and fourth degree terms to the relative bias of a ratio estimate is $3C_x^2/n$ times the value of the latter to a first approximation. Unless n is small, the contribution can, therefore, be considered to be negligible. G. R. Ayachit (1953) has assessed the value of contributions to the bias from successive approximations by means of experimental sampling on a wide range of populations commonly met with in surveys, and found that the contribution of higher order terms is negligible. For appreciably large n, say 30 or larger, even the leading term is found to be of no consequence.

4a.5 Variance of the Ratio Estimate

By definition,

$$V(R_n) = E\{R_n - E(R_n)\}^2 \tag{20}$$

From (8) we write to a first approximation

$$R_n = R_N + R_N\left(\frac{\bar{\epsilon}_n}{\bar{y}_N} - \frac{\bar{\epsilon}_n{}'}{\bar{x}_N}\right) + R_N\left(\frac{\bar{\epsilon}_n{}'^2}{\bar{x}_N{}^2} - \frac{\bar{\epsilon}_n\bar{\epsilon}_n{}'}{\bar{y}_N\bar{x}_N}\right) \tag{21}$$

Hence, substituting from (21) and (12) in (20), we have on expanding and retaining terms up to the second degree,

$$\begin{aligned} E\{R_n - E(R_n)\}^2 &= R_N{}^2E\left(\frac{\bar{\epsilon}_n}{\bar{y}_N} - \frac{\bar{\epsilon}_n{}'}{\bar{x}_N}\right)^2 \\ &= R_N{}^2E\left(\frac{\bar{\epsilon}_n{}^2}{\bar{y}_N{}^2} + \frac{\bar{\epsilon}_n{}'^2}{\bar{x}_N{}^2} - \frac{2\bar{\epsilon}_n\bar{\epsilon}_n{}'}{\bar{y}_N\bar{x}_N}\right) \end{aligned} \tag{22}$$

Let V_1 denote the variance of a ratio estimate to a first approximation. Taking the expectations term by term, we obtain

$$V_1(R_n) = R_N{}^2\,\frac{N-n}{N}\,\frac{1}{n}\left(C_y{}^2 + C_x{}^2 - 2\rho C_yC_x\right) \tag{23}$$

or

$$V_1\left(\frac{R_n}{R_N}\right) = \frac{N-n}{N}\,\frac{1}{n}\left(C_y^2 + C_x^2 - 2\rho C_y C_x\right) \tag{24}$$

When

$$C_y^2 = C_x^2 = C^2$$

the expression for the relative variance takes the form

$$V_1\left(\frac{R_n}{R_N}\right) = \frac{N-n}{Nn}\,2C^2(1-\rho) \tag{25}$$

and, for large N,

$$V_1\left(\frac{R_n}{R_N}\right) = \frac{2C^2}{n}\,(1-\rho)$$

$$= 2\ \text{(relative bias)} \tag{26}$$

Comparing (26) with (13), we see that the bias in a ratio estimate is of the order $1/n$ and hence negligible, for n sufficiently large, as compared to its standard error which is of the order $1/\sqrt{n}$.

To obtain the variance of the estimate of the total, namely, Y_R, we multiply (23) by $N^2\bar{x}_N^2$, so that

$$V_1(Y_R) = \frac{N-n}{Nn}\,Y^2\left(C_y^2 + C_x^2 - 2\rho C_y C_x\right) \tag{27}$$

which can also be alternatively written as

$$= \frac{N(N-n)}{n}\left\{\mathbf{S}_y^2 + R_N^2\mathbf{S}_x^2 - 2R_N\,\rho\mathbf{S}_y\mathbf{S}_x\right\}$$

$$= N^2\left\{V(\bar{y}_n) + R_N^2 V(\bar{x}_n) - 2R_N\,\mathrm{Cov}\,(\bar{y}_n, \bar{x}_n)\right\} \tag{28}$$

An alternative expression for the variance of the ratio estimate which, in some ways, is more instructive can be obtained by expanding the expression within brackets in (27). We have

$$C_y^2 + C_x^2 - 2\rho C_y C_x = \frac{\mathbf{S}_y^2}{\bar{y}_N^2} + \frac{\mathbf{S}_x^2}{\bar{x}_N^2} - \frac{2\rho\mathbf{S}_y\mathbf{S}_x}{\bar{y}_N\bar{x}_N}$$

$$= \frac{1}{\bar{y}_N^2}\left\{\frac{\sum\limits_{i=1}^{N} y_i^2 - N\bar{y}_N^2}{N-1} + \frac{\bar{y}_N^2}{\bar{x}_N^2}\cdot\frac{\left(\sum\limits_{i=1}^{N} x_i^2 - N\bar{x}_N^2\right)}{N-1} - \frac{2\bar{y}_N}{\bar{x}_N}\cdot\frac{\left(\sum\limits_{i=1}^{N} y_i x_i - N\bar{y}_N\bar{x}_N\right)}{N-1}\right\}$$

$$= \frac{1}{\bar{y}_N^2\,(N-1)}\left\{\sum_{i=1}^{N} y_i^2 + R_N^2 \sum_{i=1}^{N} x_i^2 - 2R_N \sum_{i=1}^{N} y_i x_i\right\}$$

$$= \frac{1}{\bar{y}_N^2\,(N-1)}\left\{\sum_{i=1}^{N}(y_i - R_N x_i)^2\right\} \tag{29}$$

Hence, from (23) and (27),

$$V_1(R_n) = \frac{N-n}{N-1}\cdot\frac{1}{n}\cdot\frac{1}{\bar{x}_N^2}\cdot\frac{1}{N}\left\{\sum_{i=1}^{N}(y_i - R_N x_i)^2\right\}. \tag{30}$$

and

$$V_1(Y_R) = \frac{N(N-n)}{(N-1)\,n}\sum_{i=1}^{N}(y_i - R_N x_i)^2 \tag{31}$$

If the population is regarded as divided into k classes with the N_i units in the i-th class having the value x_i each, (30) and (31) can be rewritten to read as follows:

$$V_1(R_n) = \frac{N-n}{N-1}\cdot\frac{1}{n}\cdot\frac{1}{\bar{x}_N^2}\cdot\frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{N_i}(y_{ij} - R_N x_i)^2 \tag{32}$$

and

$$V_1(Y_R) = \frac{N(N-n)}{(N-1)\;n}\left\{\sum_{i=1}^{k}\sum_{j=1}^{N_i}(y_{ij} - R_N x_i)^2\right\} \tag{33}$$

Clearly, the term under the first summation is proportional to the variance of y for a fixed value of x when the regression of y on x is linear, and the regression line passes through the origin. For, in this case, we have from (14) and (15),

$$E(y_{ij} \mid x_i) = R_N x_i$$

Hence

$$\sum_{j=1}^{N_i} (y_{ij} - R_N x_i)^2 = \sum_{j=1}^{N_i} \{y_{ij} - E(y_{ij} \mid x_i)\}^2$$

$$= N_i \, V(y_{ij} \mid x_i)$$

or simply

$$= N_i \, V(y \mid i)$$

We may, therefore, write (32) and (33) as follows:

$$V_1(R_n) = \frac{N-n}{N-1} \cdot \frac{1}{n} \cdot \frac{1}{\bar{x}_N{}^2} \frac{1}{N} \sum_{i=1}^{k} N_i \, V(y_{ij} \mid x_i) \tag{34}$$

and

$$V_1(Y_R) = \frac{N(N-n)}{N-1} \cdot \frac{1}{n} \sum_{i=1}^{k} N_i \; V(y_{ij} \mid x_i) \tag{35}$$

It is, therefore, seen that the variance of a ratio estimate depends upon the relationship between the variance of y for a fixed x and the value of x. The situations of practical importance are those in which

$$(a) \;\; V(y \mid x) = \text{constant say } \gamma, \quad \text{or } V\left(\frac{y}{x} \mid x\right) = \frac{\gamma}{x^2} \tag{36a}$$

$$(b) \;\; V(y \mid x) \propto x \qquad \text{say } \gamma x, \quad \text{or } V\left(\frac{y}{x} \mid x\right) = \frac{\gamma}{x} \tag{36b}$$

and

$$(c) \;\; V(y \mid x) \propto x^2 \qquad \text{say } \gamma x^2, \text{ or } V\left(\frac{y}{x} \mid x\right) = \gamma \tag{36c}$$

On substitution in (34) and (35) we obtain the approximate values given on the next page for the variance of a ratio estimate appropriate to the above three cases.

	$V_1(R_n)$	$V_1(Y_R)$	
(*a*)	$\frac{N-n}{N-1}\cdot\frac{\gamma}{n\bar{x}_N^2}$	$\frac{N^2(N-n)}{N-1}\cdot\frac{\gamma}{n}$	(37 *a*)
(*b*)	$\frac{N-n}{N-1}\cdot\frac{\gamma}{n\bar{x}_N}$	$\frac{N^2(N-n)}{N-1}\cdot\frac{\gamma}{n}\bar{x}_N$	(37 *b*)
(*c*)	$\frac{N-n}{N-1}\cdot\frac{\gamma}{n}\cdot\frac{\sum_{i=1}^{k}N_i x_i^2}{N\bar{x}_N^2}$	$\frac{N(N-n)}{N-1}\cdot\frac{\gamma}{n}\sum_{i=1}^{k}N_i x_i^2$	(37 *c*)

We will later on show that when $V(y \mid x) \propto x$, the ratio estimate is the best unbiased linear estimate for a given set of x's.

4*a*.6 Estimate of the Variance of the Ratio Estimate

Just as s_y^2, $\bar{y}_n$, s_x^2 and $\bar{x}_n$ provide unbiased estimates of the corresponding population values, similarly s_{yx} defined by

$$s_{yx} = \frac{\sum^{n}(y_i-\bar{y}_n)(x_i-\bar{x}_n)}{n-1}$$

provides an unbiased estimate of the corresponding population value $\rho\mathbf{S}_y\mathbf{S}_x$. If the sample means and variances are independently distributed as they will, for instance, be in samples of n from a normal bivariate population, an estimate of the relative variance of a ratio estimate will be given by

$$\text{Est. } V_1\left(\frac{R_n}{R_N}\right) = \frac{N-n}{Nn}\left(\frac{s_y^2}{\bar{y}_n^2}+\frac{s_x^2}{\bar{x}_n^2}-\frac{2s_{yx}}{\bar{y}_n\bar{x}_n}\right) \tag{38}$$

On simplification in the manner shown in the previous section, (38) reduces to

$$\text{Est. } V_1\left(\frac{R_n}{R_N}\right) = \frac{N-n}{Nn}\cdot\frac{1}{\bar{y}_n^2}\cdot\frac{1}{n-1}\sum^{n}(y_i-R_n x_i)^2 \tag{39}$$

We can thus take the estimates of the variances of R_n and Y_R to be

$$\text{Est. } V_1(R_n) = \frac{N-n}{Nn}\cdot\frac{1}{\bar{x}_n^2}\cdot\frac{1}{n-1}\sum^{n}(y_i-R_n x_i)^2 \tag{40}$$

and

$$\text{Est. } V_1(Y_R) = \frac{N(N-n)}{n} \cdot \frac{1}{n-1} \sum^{n} (y_i - R_n x_i)^2 \tag{41}$$

The reader will note that these are biased estimates but the bias will be negligible if the coefficients of variation of y and x are small.

One special case of a ratio estimate, for which the estimated variance takes a particularly simple form, may be mentioned. It is the case of a weighted mean in which the weights are in the nature of ancillary information, varying from one sampling unit to another, with y_i of the form $w_i\eta_i$ and $x_i = w_i$, so that

$$R_n = \bar{\eta}_w = \frac{\overset{n}{\Sigma}\, w_i\eta_i}{\overset{n}{\Sigma}\, w_i} \tag{42}$$

The sample estimate of the variance for this case is given by

$$\text{Est. } V_1(\bar{\eta}_w) = \frac{N-n}{Nn} \cdot \frac{1}{\bar{w}_n^{\,2}} \cdot \frac{1}{(n-1)} \sum^{n} w_i^2 (\eta_i - \bar{\eta}_w)^2 \tag{43}$$

and

$$\text{Est. } V_1\ (Y_R) = \frac{N(N-n)}{n} \cdot \frac{1}{(n-1)} \sum^{n} w_i^2 (\eta_i - \bar{\eta}_w)^2 \tag{44}$$

4*a*.7* Second Approximation to the Variance of the Ratio Estimate

We shall first obtain an expression for the relative mean square error of R_n defined as

$$\frac{1}{R_N^{\,2}}\, E(R_n - R_N)^2 \tag{45}$$

From (8), we have to a second approximation, *i.e.*, neglecting the contribution of terms involving powers in $\bar{\epsilon}_n$ and $\bar{\epsilon}_n'$ higher than the fourth,

$$\frac{R_n - R_N}{R_N} = \frac{\bar{\epsilon}_n}{\bar{y}_N} - \frac{\bar{\epsilon}_n'}{\bar{x}_N} + \frac{\bar{\epsilon}_n'^2}{\bar{x}_N^{\,2}} - \frac{\bar{\epsilon}_n\bar{\epsilon}_n'}{\bar{y}_N\bar{x}_N} + \frac{\bar{\epsilon}_n\bar{\epsilon}_n'^2}{\bar{y}_N\bar{x}_N^{\,2}} - \frac{\bar{\epsilon}_n'^3}{\bar{x}_N^{\,3}} + \frac{\bar{\epsilon}_n'^4}{\bar{x}_N^{\,4}} - \frac{\bar{\epsilon}_n\bar{\epsilon}_n'^3}{\bar{y}_N\bar{x}_N^{\,3}} \tag{46}$$

On squaring both sides, expanding the right-hand side and retaining terms up to and including the fourth power in $\bar{\epsilon}_n$ and $\bar{\epsilon}_n'$ and taking expectations, we obtain

$$E\left(\frac{R_n - R_N}{R_N}\right)^2 = E\left(\frac{\bar{\epsilon}_n^{\,2}}{\bar{y}_N^{\,2}} + \frac{\bar{\epsilon}_n'^{\,2}}{\bar{x}_N^{\,2}} - \frac{2\bar{\epsilon}_n\bar{\epsilon}_n'}{\bar{y}_N\bar{x}_N}\right)$$

$$+ E\left(\frac{4\bar{\epsilon}_n\bar{\epsilon}_n'^{\,2}}{\bar{y}_N\bar{x}_N^{\,2}} - \frac{2\bar{\epsilon}_n^{\,2}\bar{\epsilon}_n'}{\bar{y}_N^{\,2}\bar{x}_N} - \frac{2\bar{\epsilon}_n'^{\,3}}{\bar{x}_N^{\,3}}\right)$$

$$+ E\left(\frac{3\bar{\epsilon}_n^{\,2}\bar{\epsilon}_n'^{\,2}}{\bar{y}_N^{\,2}\bar{x}_N^{\,2}} - \frac{6\bar{\epsilon}_n\bar{\epsilon}_n'^{\,3}}{\bar{y}_N\bar{x}_N^{\,3}} + \frac{3\bar{\epsilon}_n'^{\,4}}{\bar{x}_N^{\,4}}\right) \tag{47}$$

We notice that the second approximation to the mean square error involves the addition of the last two terms in (47) to the expression as given in (22). Using the results from the Appendix, we obtain

$$E\left(\frac{R_n - R_N}{R_N}\right)^2 = \frac{N-n}{N-1}\,\frac{1}{n}\left(\frac{\mu_{20}}{\mu_{10}^{\;2}} + \frac{\mu_{02}}{\mu_{01}^{\;2}} - \frac{2\mu_{11}}{\mu_{10}\mu_{01}}\right)$$

$$+ \frac{2\,(N-n)\,(N-2n)}{(N-1)\,(N-2)}\,\frac{1}{n^2}$$

$$\times\left(\frac{2\mu_{12}}{\mu_{10}\mu_{01}^{\;2}} - \frac{\mu_{21}}{\mu_{10}^{\;2}\mu_{01}} - \frac{\mu_{03}}{\mu_{01}^{\;3}}\right)$$

$$+ \frac{3\,(N-n)\,(N^2+N-6nN+6n^2)}{n^3\,(N-1)\,(N-2)\,(N-3)}$$

$$\times\left(\frac{\mu_{22}}{\mu_{10}^{\;2}\mu_{01}^{\;2}} - \frac{2\mu_{13}}{\mu_{10}\mu_{01}^{\;3}} + \frac{\mu_{04}}{\mu_{01}^{\;4}}\right)$$

$$+ \frac{3\,(n-1)\,N\,(N-n)\,(N-n-1)}{n^3\,(N-1)\,(N-2)\,(N-3)}$$

$$\times\left(\frac{\mu_{20}\mu_{02} + 2\mu_{11}^{\;2}}{\mu_{10}^{\;2}\mu_{01}^{\;2}} - \frac{6\mu_{11}\mu_{02}}{\mu_{10}\mu_{01}^{\;3}} + \frac{3\mu_{02}^{\;2}}{\mu_{01}^{\;4}}\right) \tag{48}$$

If the population is large and follows the bivariate normal distribution, we have

$$\mu_{30} = \mu_{21} = \mu_{12} = \mu_{03} = 0$$

$$\mu_{04} = 3\mathrm{S}_x^{\;4},\ \mu_{40} = 3\mathrm{S}_y^{\;4},\ \mu_{31} = 3\rho\mathrm{S}_y^{\;3}\mathrm{S}_x,\ \mu_{13} = 3\rho\mathrm{S}_y\mathrm{S}_x^{\;3}$$

$$\mu_{22} = (1 + 2\rho^2)\ \mathrm{S}_y^{\;2}\mathrm{S}_x^{\;2}$$

so that

$$E\left(\frac{R_n - R_N}{R_N}\right)^2 = \frac{1}{n}\left\{\frac{\mathbf{S}_y^{\,2}}{\bar{y}_N^{\,2}} + \frac{\mathbf{S}_x^{\,2}}{\bar{x}_N^{\,2}} - \frac{2\rho\mathbf{S}_y\mathbf{S}_x}{\bar{y}_N\bar{x}_N}\right\}$$
$$+ \frac{3}{n^2}\left\{(1+2\rho^2)\frac{\mathbf{S}_y^{\,2}\mathbf{S}_x^{\,2}}{\bar{y}_N^{\,2}\bar{x}_N^{\,2}} - \frac{6\rho\mathbf{S}_y\mathbf{S}_x^{\,3}}{\bar{y}_N\bar{x}_N^{\,3}} + \frac{3\mathbf{S}_x^{\,4}}{\bar{x}_N^{\,4}}\right\}$$
$$= \frac{1}{n}\left\{\frac{\mathbf{S}_y^{\,2}}{\bar{y}_N^{\,2}} + \frac{\mathbf{S}_x^{\,2}}{\bar{x}_N^{\,2}} - \frac{2\rho\mathbf{S}_y\mathbf{S}_x}{\bar{y}_N\bar{x}_N}\right\}$$
$$+ \frac{3\mathbf{S}_x^{\,2}}{n^2\bar{x}_N^{\,2}}\left\{\frac{\mathbf{S}_y^{\,2}}{\bar{y}_N^{\,2}} + \frac{\mathbf{S}_x^{\,2}}{\bar{x}_N^{\,2}} - \frac{2\rho\mathbf{S}_y\mathbf{S}_x}{\bar{y}_N\bar{x}_N}\right.$$
$$\left. + \frac{2\mathbf{S}_x^{\,2}}{\bar{x}_N^{\,2}} + \frac{2\rho^2\mathbf{S}_y^{\,2}}{\bar{y}_N^{\,2}} - \frac{4\rho\mathbf{S}_y\mathbf{S}_x}{\bar{y}_N\bar{x}_N}\right\}$$

which can be expressed as

$$= V_1\left(\frac{R_n}{R_N}\right) + \frac{3}{n}C_x^{\,2}\left\{V_1\left(\frac{R_n}{R_N}\right) + \frac{2}{n}(C_x - \rho C_y)^2\right\} \tag{49}$$

where $V_1(R_n/R_N)$ denotes the relative variance of R_n to a first approximation in samples of n from a large population.

For the finite population, the effect will be approximately to multiply $V_1(R_n/R_N)$ by $(N-n)/N$. We may, therefore, write

$$E\left(\frac{R_n - R_N}{R_N}\right)^2 = V_1\left(\frac{R_n}{R_N}\right) + \frac{3}{n}C_x^{\,2}\left\{V_1\left(\frac{R_n}{R_N}\right)\right.$$
$$\left. + \frac{2}{n}(C_x - \rho C_y)^2\right\} \tag{50}$$

Now

$$E(R_n - R_N)^2 = E[R_n - E(R_n)]^2 + [E(R_n) - R_N]^2$$
$$= V(R_n) + (\text{bias})^2 \tag{51}$$

Hence, deducting from (50) the square of the relative bias term given by (13), we obtain

$$V_2\left(\frac{R_n}{R_N}\right) = V_1\left(\frac{R_n}{R_N}\right) + \frac{3}{n}C_x^{\,2}V_1\left(\frac{R_n}{R_N}\right) + \frac{5C_x^{\,2}}{n^2}(C_x - \rho C_y)^2 \tag{52}$$

When $C_x = C_y = C$, the expression simplifies to

$$V_2\left(\frac{R_n}{R_N}\right) = \frac{2C^2(1-\rho)}{n} + \frac{C^4}{n^2}\{6(1-\rho) + 5(1-\rho)^2\} \tag{53}$$

Since $(1-\rho)^2$ will usually be negligible as compared to $(1-\rho)$, we have

$$V_2\left(\frac{R_n}{R_N}\right) = V_1\left(\frac{R_n}{R_N}\right)\left\{1 + \frac{3}{n}C^2\right\} \tag{54}$$

It will be seen that the expression for the second approximation to the relative variance is related to the first approximation in the same way as the expression for the second approximation to the relative bias is related to the first. We conclude that, unless n is too small, the first approximation may be considered as adequate. The result is due to Cochran (1940).

4*a*.8 Conditions for a Ratio Estimate to be the Best Unbiased Linear Estimate

We shall now show that when (i) the relationship between the mean value of y for a given x is linear with x and passes through the origin, and (ii) the variance of y about this line is proportional to x, then, for a given set of x's, the ratio estimate Y_R gives the best unbiased linear estimate of the population total and its variance is given by

$$\frac{N-n}{N} \cdot \frac{\gamma}{n\bar{x}_n} \cdot N^2\bar{x}_N{}^2$$

Let equation (14), namely,

$$E(y \mid x) = \beta x \tag{55}$$

denote the regression of y on x passing through the origin, and

$$V(y \mid x) = \gamma x \tag{56}$$

denote the relationship between the variance of y for a given x, and x. We have seen in (15) that β in this case represents the population ratio R_N.

It is known from Section 2*a*.3 that the best unbiased linear estimate is given by a corollary of the Markoff theorem on linear estimation (Neyman and David, 1938). The method consists in setting up a linear function of observations as the estimate of β and minimizing its variance subject to the condition that the estimate is an unbiased estimate of β. Suppose that the estimate of the population total Y is given by

$$Y_R = \sum_{i=1}^{k} \lambda_i n_i \bar{y}_{n_i} \tag{57}$$

where the y observations in the same class are assumed to have equal weight and λ_i's are chosen by the application of the Markoff method. Now, the condition of unbiasedness gives

$$E\left(\sum_{i=1}^{k} \lambda_i n_i \bar{y}_{n_i}\right) = Y = \sum_{i=1}^{k} N_i \bar{y}_{N_i} \tag{58}$$

Substituting from (55), we obtain

$$\sum_{i=1}^{k} \lambda_i n_i \beta x_i = \sum_{i=1}^{k} N_i \beta x_i$$

or

$$\sum_{i=1}^{k} (n_i \lambda_i - N_i)\, x_i = 0 \tag{59}$$

Now the variance of Y_R for given $n_1, n_2, \ldots, n_k$ is given by

$$\begin{aligned} V\,(Y_R \mid n_1, n_2, \ldots, n_k) &= E\left\{\left(\sum_{i=1}^{k} \lambda_i n_i \bar{y}_{n_i} - \sum_{i=1}^{k} \lambda_i n_i \bar{y}_{N_i}\right)^2 \middle| \begin{matrix} n_1 \\ n_2 \\ \vdots \\ n_k \end{matrix} \right\} \\ &= E\left\{\sum_{i=1}^{k} \lambda_i^2 n_i^2 (\bar{y}_{n_i} - \bar{y}_{N_i})^2 \right. \\ &\qquad \left. + \sum_{i \neq j=1}^{k} \lambda_i \lambda_j n_i n_j (\bar{y}_{n_i} - \bar{y}_{N_i})(\bar{y}_{n_j} - \bar{y}_{N_j})\right\} \\ &= \sum_{i=1}^{k} n_i^2 \lambda_i^2\, V(\bar{y}_{n_i} \mid x_i) \end{aligned} \tag{60}$$

From (56), we have

$$V(y \mid x = x_i) = \gamma x_i$$

i.e.,

$$\frac{N_i - 1}{N_i} \mathbf{S}_i^2 = \gamma x_i$$

where $\mathbf{S}_i^2$ is the mean square for the i-th class. Hence

$$\begin{aligned} V(\bar{y}_{n_i} \mid x_i) &= \frac{N_i - n_i}{N_i} \cdot \frac{1}{n_i} \mathbf{S}_i^2 \\ &= \frac{N_i - n_i}{N_i} \cdot \frac{1}{n_i} \frac{N_i}{N_i - 1} \gamma x_i \\ &= \frac{N_i - n_i}{N_i - 1} \frac{1}{n_i} \gamma x_i \qquad (61) \end{aligned}$$

Substituting in (60), we get

$$V(Y_R \mid n_1, n_2, \ldots, n_k) = \gamma \sum_{i=1}^{k} \lambda_i^2 n_i \frac{N_i - n_i}{N_i - 1} x_i \qquad (62)$$

The Markoff method of estimation requires that λ_i's are to be so determined that (62) is minimum subject to the condition that (59) holds.

Let

$$\phi = \gamma \sum_{i=1}^{k} \lambda_i^2 n_i \frac{N_i - n_i}{N_i - 1} x_i - \mu \sum_{i=1}^{k} (n_i \lambda_i - N_i) x_i \qquad (63)$$

where μ is some constant. Equation (63) can be written as

$$\phi = \sum_{i=1}^{k} n_i x_i \left\{ \lambda_i \sqrt{\frac{\gamma (N_i - n_i)}{N_i - 1}} - \frac{\mu}{2} \sqrt{\frac{N_i - 1}{\gamma (N_i - n_i)}} \right\}^2$$

$$+ \text{ terms independent of } \lambda_i \qquad (64)$$

Clearly, $V(Y_R \mid n_1, n_2, \ldots, n_k)$ is minimum when each of the square terms on the right-hand side of (64) is zero, or in other words, when

$$\lambda_i = \frac{\mu}{2\gamma} \cdot \frac{N_i - 1}{N_i - n_i} \qquad (i = 1, 2, \ldots, k) \qquad (65)$$

To evaluate μ, we substitute for λ_i from (65) in (59) and obtain

$$\frac{\mu}{2\gamma} = \frac{N\bar{x}_N}{\sum_{i=1}^{k} \frac{(N_i - 1)\, n_i x_i}{(N_i - n_i)}}$$

whence

$$\lambda_i = N\bar{x}_N \frac{\left(\frac{N_i - 1}{N_i - n_i}\right)}{\sum_{i=1}^{k} \frac{N_i - 1}{N_i - n_i} n_i x_i} \qquad (i = 1, 2, \ldots, k) \tag{66}$$

Hence, substituting for λ_i from (66) in (57) and (62), we obtain

$$Y_R = \frac{N\bar{x}_N \sum_{i=1}^{k} n_i \bar{y}_{n_i} \left(\frac{N_i - 1}{N_i - n_i}\right)}{\sum_{i=1}^{k} n_i x_i \frac{N_i - 1}{N_i - n_i}} \tag{67}$$

and

$$V(Y_R \mid n_1, n_2, \ldots, n_k) = \frac{N^2 \bar{x}_N{}^2 \gamma}{\sum_{i=1}^{k} n_i x_i \left(\frac{N_i - 1}{N_i - n_i}\right)} \tag{68}$$

If the sampling in each class is assumed to be carried out with replacement, or alternatively, n_i is small compared with N_i so that $(N_i - 1)/(N_i - n_i)$ can be assumed to be unity, (67) and (68) are seen to reduce to

$$Y_R = N\bar{x}_N \left(\frac{\bar{y}_n}{\bar{x}_n}\right) \tag{69}$$

and

$$V(Y_R \mid n_1, n_2, \ldots, n_k) = \gamma \frac{N^2 \bar{x}_N{}^2}{n\bar{x}_n} \tag{70}$$

showing that the ratio estimate under the conditions stated in the beginning of this section gives the best unbiased linear estimate, provided N_i's are large.

When, however, N_i's are not large, the estimate Y_R will be only approximately given by (69) and the effect on the variance of Y_R will be to multiply (70) by the usual finite multiplier $(N-n)/N$. For, estimating N_i from the sample, we have

$$N_i = n_i \left(\frac{N}{n}\right)$$

and hence

$$\frac{N_i - 1}{N_i - n_i} \cong \frac{N}{N-n} \tag{71}$$

On substituting from (71) in (67) and (68), we obtain

$$Y_R = N\bar{x}_N \cdot R_n \tag{72}$$

and

$$V(Y_R \mid n_1, n_2, \ldots, n_k) = \frac{N-n}{N} \cdot \gamma \cdot \frac{N^2\bar{x}_N{}^2}{n\bar{x}_n} \tag{73}$$

We notice that the variance depends upon the set of x's which happen to turn up in the sample. In repeated samples, to a first approximation, the average value of (73) is given by

$$V(Y_R) = \frac{N(N-n)}{n} \gamma\bar{x}_N \tag{74}$$

The slight difference between this expression and (37*b*) is due to the use of approximations in the derivation of both.

4*a*.9 Confidence Limits

We have just seen that when (i) the relationship between y and x is a straight line passing through the origin, and (ii) the variance of y about this line is proportional to x, the ratio estimate R_n is the best unbiased linear estimate of the population ratio for a given set of x's, with the sampling variance given by

$$\frac{N-n}{N} \cdot \frac{\gamma}{n\bar{x}_N}$$

It is a well-known property of an unbiased linear estimate that if n is not too small and N is large, the probability that the difference between the estimate and the population value will exceed a fixed multiple of the standard error of the estimate is approximately equal to the probability as determined by the normal law. Consequently, the confidence limits for a population ratio are obtained in the manner indicated in Chapter II, being given by

$$R_n \pm t_{(\alpha,\,\infty)} \sqrt{\frac{\gamma}{n\bar{x}_N}} \tag{75}$$

When, however, conditions (i) and (ii) of Section $4a.8$ do not hold, the exact distribution of a ratio estimate is not known to have been expressed in a simple form. For large samples, however, the distribution of R_n/R_N can be regarded as normal for all practical purposes, with the standard error given by

$$\frac{1}{\sqrt{n}}\,(C_x^2 - 2\rho C_x C_y + C_y^2)^{\frac{1}{2}}$$

We, therefore, expect the following inequality to hold on the average with probability $(1 - \alpha)$:

$$\frac{R_n}{R_N} - t_{(\alpha,\,\infty)} \frac{1}{\sqrt{n}}\,(C_x^2 - 2\rho C_x C_y + C_y^2)^{\frac{1}{2}} \leqslant 1 \leqslant \frac{R_n}{R_N} + t_{(\alpha,\,\infty)} \frac{1}{\sqrt{n}}\,(C_x^2 - 2\rho C_x C_y + C_y^2)^{\frac{1}{2}}$$

yielding the following confidence limits for R_N:

$$\frac{R_n}{1 \pm t_{(\alpha,\,\infty)} \cdot \frac{1}{\sqrt{n}}\,(C_x^2 - 2\rho C_x C_y + C_y^2)^{\frac{1}{2}}} \tag{76}$$

For small samples, the following method is available. Let (x_i, y_i) be normally distributed. Consider a function

$$u = \bar{y}_n - R_N \bar{x}_n \tag{77}$$

Clearly, u will be normally distributed with variance

$$\mathbf{S}_u^2 = \frac{1}{n}\,(\mathbf{S}_y^2 - 2R_N \rho \mathbf{S}_y \mathbf{S}_x + R_N^2 \mathbf{S}_x^2) \tag{78}$$

The confidence limits for R_N with the confidence coefficient $(1 - \alpha)$ are then determined by the two roots of the quadratic in R_N given by

$$t_{(\alpha,\, \infty)} = \frac{\sqrt{n}\,(\bar{y}_n - R_N\bar{x}_n)}{(\mathbf{S}_y{}^2 - 2R_N\rho\mathbf{S}_y\mathbf{S}_x + R_N{}^2\mathbf{S}_x{}^2)^{\frac{1}{2}}} \tag{79}$$

Solving the above quadratic, we obtain for the confidence limits of R_N the following values:

$$\frac{R_n}{\left\{1 - \dfrac{t^2{}_{(\alpha,\, \infty)}\, C_x{}^2}{n}\right\}} \left[\left(1 - \frac{t^2{}_{(\alpha,\, \infty)}}{n}\, \rho C_y C_x\right) \pm \frac{t_{(\alpha,\, \infty)}}{\sqrt{n}} \left\{(C_x{}^2 - 2\rho C_x C_y + C_y{}^2) - \frac{t^2{}_{(\alpha,\, \infty)}}{n}\, C_x{}^2 C_y{}^2\, (1 - \rho^2)\right\}^{\frac{1}{2}}\right] \tag{80}$$

where

$$C_x = \frac{\mathbf{S}_x}{\bar{x}_n}, \quad \text{and} \quad C_y = \frac{\mathbf{S}_y}{\bar{y}_n} \tag{81}$$

4*a*.10 Efficiency of the Ratio Estimate

We have seen that the variance of the estimate of the population total based on the simple arithmetic mean is given by

$$N\,(N - n)\, \frac{\mathbf{S}_y{}^2}{n}$$

We also saw that the first approximation to the variance of the estimate of the population total based on the ratio method is given by

$$\frac{N\,(N - n)}{n}\, \{\mathbf{S}_y{}^2 + R_N{}^2\mathbf{S}_x{}^2 - 2R_N\rho\mathbf{S}_y\mathbf{S}_x\}$$

Now the relative efficiency of an estimate B compared to that of another estimate A based on a sample of equal size is defined in Section 3*a*.10 as the ratio of the inverse of their variances. Hence

$$\text{Efficiency} = \frac{S_y^2}{S_y^2 + R_N^2 S_x^2 - 2R_N \rho S_y S_x}$$

$$= \frac{1}{1 + \left(\frac{C_x^2}{C_y^2}\right) - 2\rho\left(\frac{C_x}{C_y}\right)} \qquad (82)$$

It follows that, in large samples, the ratio estimate will be more efficient than the corresponding sample estimate based on the simple arithmetic mean if the denominator is less than 1, *i.e.*, if

$$\left(\frac{C_x^2}{C_y^2}\right) < 2\rho\left(\frac{C_x}{C_y}\right)$$

or

$$\rho > \tfrac{1}{2}\,\frac{C_x}{C_y} \qquad (83)$$

If $C_x = C_y$, as will be the case, for example, when y and x denote values in two consecutive periods of the same variate, ρ will have to be larger than one-half in order that the ratio estimate may be more efficient than the one based on the simple arithmetic mean.

Example 4.1

A sample survey for the estimation of livestock numbers was carried out in Etawah (India) during the spring of 1951. Table 4.1 summarizes the data in respect of the number of livestock y and the agricultural area x in all the 364 villages in the Etawah subdivision. The range of agricultural area is divided into 7 classes: 0–100, 101–200, 201–300, 301–400, 401–600, 601–1000 and greater than 1000 acres; and for each of these classes, the number of villages (N_i), the mean agricultural area per village ($\bar{x}_{N_i}$), the mean number of livestock per village ($\bar{y}_{N_i}$) and the values of $V(x \mid i)$, $V(y \mid i)$ and ρ_i are given ($i = 1, 2, \ldots, 7$). The values of the mean, the variance and the coefficient of correlation obtained by combining together the first six classes, *i.e.*, grouping together all the villages having agricultural area up to 1000 acres as well as those obtained by combining together all the seven classes are also given in cols. 8 and 9 of Table 4.1. Examine, graphically or otherwise, whether conditions (i) and (ii) in Section 4*a*.8 may be considered to be satisfied, so that advantage may be taken of the ratio method of estimation for estimating the livestock population.

We have worked out in Table 4.1 the ratios of $\bar{y}_{N_i}$ to $\bar{x}_{N_i}$ and of $V(y \mid i)$ to $\bar{x}_{N_i}$ for the different classes. The ratio of $\bar{y}_{N_i}$ to $\bar{x}_{N_i}$ will be seen to be fairly constant, showing that the relationship between y and x is approximately linear. $V(y \mid i)$ also appears to vary as x up to 1000 acres but not beyond it. It has, however, to be observed that the coefficient of correlation for the last class, namely, with villages having area larger than 1000 acres, is rather large and the calculated value $V(y \mid i)$ cannot possibly give for this class a correct idea of the variance of y about the line $y = R_N x$. On the other hand, any further division of this class to study the behaviour of the variance of y with x is also not feasible owing to the fact that the number of observations is few. As about 35% of the livestock population is accounted for by villages with agricultural area larger than 1000 acres, it appears advisable to study separately areas less than 1000 acres and those with larger acreage. The ratio estimate may be used to provide an efficient estimate of the livestock population for the first group comprising all the villages in the first six classes.

Example 4.2

Calculate for a sample of 64 villages, the sampling variance of the estimate of the total livestock population for villages with area less than 1000 acres based on (*a*) the simple arithmetic mean, and (*b*) the ratio method. Hence calculate the relative efficiency of the latter as compared with the former.

For obtaining the variance of the estimate based on the simple arithmetic mean, we need the value of $V(y)$ based on all the 319 observations (N) with area below 1000 acres. This is given in col. 8 of Table 4.1.

Substituting in the formula for the variance of the estimated total based on the simple arithmetic mean, we obtain

$$V(N\bar{y}_n) = N^2 \cdot \frac{N-n}{N-1} \cdot \frac{V(y)}{n}$$

$$= 319^2 \times \frac{(319-64)}{318} \times \frac{8292}{64}$$

$$= 10572000$$

Since the sample size is fairly large, the sampling variance of the ratio estimate of the total livestock population may be assumed to be given by

$$\frac{N^2\,(N-n)}{(N-1)\,n}\,\{V(y) - 2R_N\,\rho\,\sqrt{V(x)\cdot V(y)} + R_N{}^2 V(x)\}$$

where $V(x)$, $V(y)$ and $\rho\,\sqrt{V(x)\cdot V(y)}$ are the variance of x, the variance of y and the covariance of x and y, and

$$R_N = \frac{\sum\limits_{i=1}^{6} N_i \bar{y}_{N_i}}{\sum\limits_{i=1}^{6} N_i \bar{x}_{N_i}}$$

Now, from Table 4.1, we get

$$R_N = \frac{113\cdot4}{367\cdot5} = 0\cdot3086$$

and, therefore,

$$R_N{}^2 = 0\cdot09523$$

Also

$$V(x) = 39528$$

$$V(y) = 8292$$

and

$$\rho\,\sqrt{V(x)\cdot V(y)} = 12378$$

Substituting these values in the expression for the variance of the ratio estimate of the total livestock population given above, we obtain

$$\begin{aligned} V(Y_R) &= \frac{319 \times 255}{64} \times \frac{319}{318} \times [8292 - 2 \times 0\cdot3086 \times 12378 \\ &\qquad + 0\cdot09523 \times 39528] \\ &= 1271\cdot0 \times \frac{319}{318}\,[8292 - 2 \times 3820 + 3764] \\ &= 1271\cdot0 \times 4430 \\ &= 5631000 \end{aligned}$$

Alternatively, we can consider

$$\frac{1}{N-1}\sum_{i=1}^{6} N_i V(y \mid i)$$

as an estimate of the mean square deviation from the ratio line since the correlation coefficient between the agricultural area in a village and the livestock population in each of the six classes is very low and not significant. In this case, we get the variance of the total livestock population estimated by the ratio method as

$$N(N-n) \times \frac{1}{n} \times \frac{1}{N-1}\sum_{i=1}^{6} N_i V(y \mid i)$$

$$= 319 \times 255 \times \frac{1}{64} \times 4602$$

$$= 5849000$$

This value is only slightly larger than the one calculated above, as one would expect owing to the small values of ρ within the classes and the close linear relationship between y and x.

It will be seen that the variance of the simple arithmetic mean estimate of the total livestock population exceeds the variance of the ratio estimate by 88%, showing thereby that the latter is 88% more efficient than the former. This large gain in efficiency of the ratio estimate is to be expected in view of the high correlation between the agricultural area in a village and the number of livestock in it.

TABLE 4.1*

Summary of the Data in Respect of the Number of Livestock (y) and the Agricultural Area (x) in Villages of Etawah Subdivision

Class Intervals (Agricultural Area of a Village—Acres)	0–100	101–200	201–300	301–400	401–600	601–1000	1000–	Classes (1) to (6) Combined	Classes (1) to (7) Combined
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Number of villages (N_i)	11	48	84	60	77	39	45	319	364
Mean agricultural area per village ($\bar{x}_{N_i}$)	63·7	155·3	245·7	344·4	491·6	767·5	1604·0	367·5	520·4
Mean number of livestock per village ($\bar{y}_{N_i}$)	25·4	50·1	76·0	99·2	150·8	244·4	425·1	113·4	151·9
$V(x \mid i) = \frac{N_i-1}{N_i} S_{xi}^2$	454	765	885	707	2775	8928	377255	39528	246891
$V(y \mid i) = \frac{N_i-1}{N_i} S_{yi}^2$	641	1201	3174	4055	5912	11119	58402	8292	25023
ρ_i	0·1703	0·1163	0·0562	0·1552	0·2421	0·2665	0·7254	0·6837	0·8388
$R_{N_i} = \bar{y}_{N_i}/\bar{x}_{N_i}$	0·399	0·323	0·309	0·288	0·307	0·318	0·265		
$\lambda_i = V(y \mid i)/\bar{x}_{N_i}$	10·1	7·7	12·9	11·8	12·0	14·5	36·4		

* Reproduced from an unpublished report on "Improvement of Livestock Statistics," Etawah (India), by the Indian Council of Agricultural Research, New Delhi.

4*a*.11 Ratio Estimate in Stratified Sampling

Let R_{n_t} denote the estimate of the population ratio for the t-th stratum and Y_{R_t} denote the ratio estimate of the population total of y in the t-th stratum. Then, clearly, the estimate of the population total of y over all the strata is given by

$$\begin{aligned} Y_R &= \sum_{t=1}^{k} Y_{R_t} \\ &= \sum_{t=1}^{k} \left\{ R_{n_t} \times \sum_{i=1}^{N_t} x_{ti} \right\} \\ &= \sum_{t=1}^{k} R_{n_t} \times N_t \bar{x}_{N_t} \qquad (84) \end{aligned}$$

Now

$$E(Y_R) = \sum_{t=1}^{k} N_t \bar{x}_{N_t} E(R_{n_t}) \qquad (85)$$

Hence, from (12), we have

$$E(Y_R) = \sum_{t=1}^{k} N_t \bar{y}_{N_t} \left\{ 1 + \frac{N_t - n_t}{N_t n_t} \left(\frac{S_{tx}{}^2}{\bar{x}_{N_t}{}^2} - \rho_t \frac{S_{ty} S_{tx}}{\bar{y}_{N_t} \bar{x}_{N_t}} \right) \right\} \qquad (86)$$

It follows that Y_R is a biased but consistent estimate. To obtain an idea of how the bias diminishes with the size of the sample, we will suppose that the finite multiplier approximates to unity, $n_t = n/k$ and further assume that $S_{tx}/\bar{x}_{N_t}$, $S_{ty}/\bar{y}_{N_t}$ and ρ_t are each of the same order from stratum to stratum, say C_x, C_y and ρ respectively. The relative bias in the estimate then equals

$$\frac{k}{n} (C_x{}^2 - \rho C_y C_x)$$

It follows that in order that Y_R should provide a satisfactory estimate of the population total, the sample size within each stratum should be sufficiently large.

The variance of Y_R, to a first approximation, is given by

$$\begin{aligned} V_1(Y_R) &= E(Y_R - Y)^2 \\ &= E\left\{\sum_{t=1}^{k} (R_{n_t}N_t\bar{x}_{N_t} - R_{N_t}N_t\bar{x}_{N_t})\right\}^2 \\ &= E\left\{\sum_{t=1}^{k} N_t^2\bar{x}_{N_t}^2(R_{n_t} - R_{N_t})^2 + \sum_{t\neq t'=1}^{k} N_tN_{t'}\bar{x}_{N_t}\bar{x}_{N_{t'}} \right. \\ &\qquad \left. \times (R_{n_t} - R_{N_t})(R_{n_{t'}} - R_{N_{t'}})\right\} \end{aligned} \tag{87}$$

Since to a first approximation

$$E(R_{n_t}) = R_{N_t}$$

and sampling is done independently in the different strata, the product term is zero and we obtain

$$V_1(Y_R) = \sum_{t=1}^{k} N_t^2\bar{x}_{N_t}^2V_1(R_{n_t}) \tag{88}$$

$$= \sum_{t=1}^{k} N_t^2\bar{x}_{N_t}^2R_{N_t}^2 \; \frac{N_t - n_t}{N_tn_t}\left(\frac{S_{ty}^2}{\bar{y}_{N_t}^2} + \frac{S_{tx}^2}{\bar{x}_{N_t}^2} - \frac{2\rho_tS_{ty}S_{tx}}{\bar{y}_{N_t}\bar{x}_{N_t}}\right)$$

$$= N\sum_{t=1}^{k}\frac{p_t(N_t - n_t)}{n_t}(S_{ty}^2 + R_{N_t}^2S_{tx}^2 - 2R_{N_t}\rho_tS_{ty}S_{tx}) \tag{89}$$

where $p_t = N_t/N$. Using (29), this can also be written as

$$V_1(Y_R) = N\sum_{t=1}^{k}\frac{p_t(N_t - n_t)}{n_t} \cdot \frac{1}{N_t - 1}\left\{\sum_{i=1}^{N_t}(y_{ti} - R_{N_t}x_{ti})^2\right\} \tag{90}$$

The above formulæ are based on the assumption that n_t is large. This, however, is not always true in practice. To get over this difficulty, Hansen, Hurwitz and Gurney (1946) suggest a single combined ratio, namely,

$$\frac{\sum_{t=1}^{k} p_t\bar{y}_{n_t}}{\sum_{t=1}^{k} p_t\bar{x}_{n_t}} \tag{91}$$

and denote this ratio by R_{n_c} and the estimate of the population total by Y_{R_c} in order to distinguish them from the corresponding estimates based on separate strata. To obtain the expected value of (91), we write

$$\sum_{t=1}^{k} p_t\bar{y}_{n_t} = \bar{y}_N + \bar{\epsilon}_n \quad \text{and} \quad \sum_{t=1}^{k} p_t\bar{x}_{n_t} = \bar{x}_N + \bar{\epsilon}_n{}' \tag{92}$$

where

$$E(\bar{\epsilon}_n) = 0, \; E(\bar{\epsilon}_n{}') = 0 \tag{93}$$

$$E(\bar{\epsilon}_n{}^2) = \sum_{t=1}^{k} \frac{N_t - n_t}{N_t n_t} p_t{}^2 \mathrm{S}_{ty}{}^2, \; E(\bar{\epsilon}_n{}'^2) = \sum_{t=1}^{k} \frac{N_t - n_t}{N_t n_t} p_t{}^2 \mathrm{S}_{tx}{}^2 \tag{94}$$

Then to a first approximation

$$E(R_{n_c}) = \frac{\bar{y}_N}{\bar{x}_N}\left\{1 + \frac{E(\bar{\epsilon}_n{}'^2)}{\bar{x}_N{}^2} - \frac{E(\bar{\epsilon}_n\bar{\epsilon}_n{}')}{\bar{y}_N\bar{x}_N}\right\} \tag{95}$$

Hence, the relative bias in R_{n_c} is given by

$$\sum_{t=1}^{k} \frac{N_t - n_t}{N_t n_t} \cdot p_t{}^2 \left(\frac{\mathrm{S}_{tx}{}^2}{\bar{x}_N{}^2} - \frac{\rho_t \mathrm{S}_{ty}\mathrm{S}_{tx}}{\bar{y}_N\bar{x}_N}\right) \tag{96}$$

To have an idea as to how rapidly the bias diminishes with the size of the sample, we will suppose that n_t is proportional to N_t and S_{tx}, S_{ty} and ρ_t are constant. The relative bias will then be seen to be given by

$$\frac{N-n}{N}\,\frac{1}{n}\,(C_x{}^2 - \rho C_y C_x) \tag{97}$$

It follows that even when the size of the sample within each stratum is small, a combined ratio estimate can give a satisfactory estimate of the population total provided the total sample is sufficiently large.

To work out the sampling variance of the combined ratio, we have, to a first approximation

$$V_1\ (Y_{Rc}) = Y^2 \left(\frac{E\,(\bar{\epsilon}_n^{\,2})}{\bar{y}_N^{\,2}} + \frac{E\,(\bar{\epsilon}_n'^{\,2})}{\bar{x}_N^{\,2}} - \frac{2E\,(\bar{\epsilon}_n \bar{\epsilon}_n')}{\bar{y}_N \bar{x}_N} \right)$$

$$= N^2 \bar{y}_N^{\,2} \left\{ \sum_{t=1}^{k} \frac{N_t - n_t}{N_t n_t} \cdot p_t^2 \cdot \frac{\mathrm{S}_{ty}^{\,2}}{\bar{y}_N^{\,2}} \right.$$

$$+ \sum_{t=1}^{k} \frac{N_t - n_t}{N_t n_t} \cdot p_t^2 \cdot \frac{\mathrm{S}_{tx}^{\,2}}{\bar{x}_N^{\,2}}$$

$$\left. - 2 \sum_{t=1}^{k} \frac{N_t - n_t}{N_t n_t} \cdot p_t^2 \cdot \frac{\rho_t \mathrm{S}_{ty} \mathrm{S}_{tx}}{\bar{y}_N \bar{x}_N} \right\}$$

$$= N \sum_{t=1}^{k} \frac{N_t - n_t}{n_t} \cdot p_t \{\mathrm{S}_{ty}^{\,2} + R_N^2 \mathrm{S}_{tx}^{\,2} - 2 R_N \rho_t \mathrm{S}_{ty} \mathrm{S}_{tx}\} \tag{98}$$

It is interesting to note that the sampling variance of the combined ratio has the same form as that of the ratio estimate based on separate strata except that there is now a single ratio R_N in place of R_{N_t}.

The difference between the sampling variance of Y_{Rc} in (98) and of Y_R in (89) is given by

$$\sum_{t=1}^{k} \frac{N p_t\,(N_t - n_t)}{n_t} \{\mathrm{S}_{tx}^{\,2}\,(R_N^2 - R_{Nt}^2) - 2\,(R_N - R_{Nt})\,\rho_t \mathrm{S}_{ty} \mathrm{S}_{tx}\}$$

$$= \sum_{t=1}^{k} \frac{N p_t\,(N_t - n_t)}{n_t} \{\mathrm{S}_{tx}^{\,2}\,(R_N - R_{Nt})^2 + 2\ \ (R_N - R_{Nt}) \times (R_{Nt} \mathrm{S}_{tx}^{\,2} - \rho_t \mathrm{S}_{ty} \mathrm{S}_{tx})\} \tag{99}$$

It will be seen that (99) depends upon the magnitude of the variation between the strata ratios and the value of

$$(R_{Nt} \mathrm{S}_{tx}^{\,2} - \rho_t \mathrm{S}_{ty} \mathrm{S}_{tx})$$

The latter will, however, be usually small, vanishing in fact when the regression of y on x is a straight line through the origin within each stratum. It follows, therefore, that the combined estimate will have a lower precision than that based on separate strata. On the other hand, the bias in the former estimate will be smaller than in the latter. Unless, therefore, the population ratios in the different strata vary considerably, the use of a combined ratio may provide an estimate which has a negligible bias and whose precision is almost as high as that of the estimate based on separate ratios.

Lastly, we shall determine the optimum allocation of the sample among the different strata when a ratio estimate is used. We shall consider the simplest case for which the cost of the survey is proportional to the size of the sample. Assuming that the cost of the survey is fixed at, say C_0, and that $C_0 = cn$, where c is the cost per unit in the sample, the optimum allocation is given by minimizing the variance of the ratio estimate given by (90) for fixed n, say n_0.

Let

$$\phi = N \sum_{t=1}^{k} \left\{ p_t \cdot \frac{N_t - n_t}{n_t} \cdot S_{ty}'^2 \right\} + \mu \left(\sum_{t=1}^{k} n_t \right) \tag{100}$$

where μ is a constant, and

$$S_{ty}'^2 = \frac{\sum_{i=1}^{N_t} (y_{ti} - R_{N_t} x_{ti})^2}{N_t - 1} \tag{101}$$

Clearly, ϕ can be written as

$$\phi = \sum_{t=1}^{k} \left\{ \frac{N_t S_{ty}'}{\sqrt{n_t}} - \sqrt{\mu n_t} \right\}^2 + \text{terms independent of } n_t \tag{102}$$

It follows that the optimum value of n_t is given by

$$n_t \propto N_t S_{ty}' \tag{103}$$

This result is thus analogous to that for the simple arithmetic mean estimate, except that instead of the variance of y within a stratum, *viz.*, S_{ty}^2, we now have the residual variance of y about the ratio line in the stratum, *viz.*, $S_{ty}'^2$.

Example 4.3

From the livestock data referred to earlier, it is proposed to draw a stratified random sample of 73 villages (amounting to 20% of the total) and to estimate the total livestock population for the entire subdivision. The villages having agricultural area up to 1000 acres constitute the first stratum and the remaining villages the second stratum. Calculate the optimum allocation of the villages between the two strata if the method of estimation to be adopted is (i) ratio method with a common ratio for both strata, (ii) ratio method with separate ratios for the two strata, and (iii) simple estimation within each stratum.

Also calculate the sampling variance of the estimated total by each of the above methods and hence compare their efficiencies.

The relevant calculations for each method step by step are presented in Tables 4.2, 4.3 and 4.4. The tables are self-explanatory. The results are tabulated below:

Method of Estimation	Number of Villages in the Sample		Sampling Variance	Efficiency
	Stratum 1	Stratum 2		
(i) Ratio estimate with common ratio for both strata	54	19	8707000	191·5
(ii) Ratio estimate with separate ratios for the two strata	54	19	8688000	192·0
(iii) Simple estimate within each stratum	53	20	16677000	100·0

TABLE 4.2

Stratified Random Sampling with a Single Combined Ratio for Both Strata

(Adopting Optimum Allocation Between the Strata)

$R_{N_c} = \frac{151\cdot9}{520\cdot4} = 0\cdot2919$ $R_{N_c}{}^2 = 0\cdot08521$

	Stratum 1 Agricultural Area 0–1000 Acres	Stratum 2 Agricultural Area > 1000 Acres
(1) N_t	319	45
(2) $V(y \mid i)$	8292	58402
(3) $\rho_i\sqrt{V(x \mid i)\,V(y \mid i)}$	12378	107673
(4) $R_{N_c}\cdot(3)$	3613	31430
(5) $V(x \mid i)$	39528	377255
(6) $R_{N_c}{}^2\cdot(5)$	3368	32146
(7) Residual *M.S.* $S_{ty}'^2 = \frac{N_t}{N_t - 1}[(2) - 2\cdot(4) + (6)]$	4448	28317
(8) $S_{ty}' = \sqrt{(7)}$	66·7	168·3
(9) $N_tS_{ty}' = (1)\cdot(8)$	21300	7600
(10) n_t*	54	19
(11) $N_t(N_t - n_t)$	84535	1170
(12) $V(Y_{R_t}) = \frac{(7)\cdot(11)}{(10)}$	6963000	1744000

$$V(Y_{R_c}) = 6963000 + 1744000$$
$$= 8707000$$
$$\therefore \ S.E. = 2951$$
$$Y = (151\cdot9)\,(364)$$
$$= 55300$$
$$\therefore \ \%S.E. = \frac{2951}{55300}\cdot100 = 5\cdot3$$

* Obtained by distributing 73, the total number of villages to be sampled, in proportion to N_tS_{ty}' given in row (9).

TABLE 4.3

Stratified Random Sampling with Separate Ratios for the Two Strata

(*Adopting Optimum Allocation Between the Strata*)

	Stratum 1 Agricultural Area 0–1000 Acres	Stratum 2 Agricultural Area >1000 Acres
(1) N_i	319	45
(2) $R_{N_i} = \bar{y}_{N_i} / \bar{x}_{N_i}$	0·3086	0·2650
(3) $R_{N_i}^2$	0·09523	0·07023
(4) $V(y \mid i)$	8292	58402
(5) $\rho_i \sqrt{V(x \mid i) \cdot V(y \mid i)}$	12378	107673
(6) (2)·(5)	3820	28533
(7) $V(x \mid i)$	39528	377255
(8) $R_{N_i}^2 V(x \mid i) = (3) \cdot (7)$	3764	26495
(9) Residual $M.S.\ S_{iy}'^2 = \frac{N_i}{N_i-1}\{(4) - 2\cdot(6) + (8)\}$	4430*	28464
(10) S_{iy}'	66·6	168·7
(11) $N_i S_{iy}' = (1)\cdot(10)$	21200	7600
(12) n_i†	54	19
(13) $N_i (N_i - n_i)$	84535	1170
(14) $V(Y_{R_i}) = \frac{(9) \cdot (13)}{(12)}$	6935000	1753000

$$V(Y_R) = 6935000 + 1753000$$

$$= 8688000$$

$$\therefore \quad S.E. = 2948$$

$$Y = (151 \cdot 9)\ (364)$$

$$= 55300$$

$$\therefore \quad \%\ S.E. = \frac{2948}{55300} \cdot 100 = 5 \cdot 3$$

* The steps leading to this figure are reproduced from example 4.2.

† Obtained by distributing the total sample in proportion to $N_i S_{iy}'$ shown in row (11).

TABLE 4.4

Stratified Random Sampling with Simple Estimate for Each Stratum

(Adopting Optimum Allocation Between the Strata)

	Stratum 1 Agricultural Area 0-1000 Acres	Stratum 2 Agricultural Area > 1000 Acres
(1) N_i	319	45
(2) $\frac{N_i-1}{N_i}\ S_{iy}^2$	8292	58402
(3) S_{iy}^2	8318	59729
(4) S_{iy}	91·2	244·4
(5) N_iS_{iy}	29100	11000
(6) n_i*	53	20
(7) $N_i\,(N_i-n_i)$	84854	1125
(8) $V\,(Y_{Ri}) = \frac{(7)\cdot(3)}{(6)}$	13317000	3360000

$$V(Y_R) = 16677000$$
$$Y = 55300$$
$$S.E. = 4084$$
$$\%\ S.E. = \frac{4084}{55300}\cdot 100 = 7{\cdot}4$$

* Obtained by distributing the total sample in proportion to N_iS_{iy} shown in row (5).

The variance of the estimate (ii) is less than that of the estimate (i) as we should expect but only slightly so. That there is no appreciable gain in assuming separate ratio lines for the two strata is also borne out by the fact that the optimum distribution of the villages in both cases turns out to be the same. Compared to the simple mean estimate, however, the ratio method is found to be considerably more efficient.

4a.12 Ratio Method for Qualitative Characters: Two Classes

We shall now consider one important application of the preceding theory to the case of qualitative characters. Suppose the population is divided into two mutually exclusive classes with N_1 and N_2 observations respectively, so that

$$N_1 = Np,\ N_2 = Nq$$

and

$$N_1 + N_2 = N$$

Assume that a simple random sample of n is chosen from N, and that n_1 of the observations in the sample are in class 1 and n_2 in class 2. We shall consider the problem of evaluating the expected value and the variance of the ratio R_n, defined by

$$R_n = \frac{n_1}{n_2} \tag{104}$$

Let y_i $(i = 1, 2, \ldots, N)$ be assumed to have the value 1 whenever it falls in class 1, and 0 if it falls in class 2; and let x_i $(i = 1, 2, \ldots, N)$ be assumed to take the value 0 whenever it falls in class 1, and 1 if it falls in class 2. It is then easy to see that

$$R_n = \frac{\sum^n y_i}{\sum^n x_i} = \frac{n_1}{n_2} \tag{105}$$

and

$$R_N = \frac{\sum_{i=1}^N y_i}{\sum_{i=1}^N x_i} = \frac{N_1}{N_2} \tag{106}$$

It follows that the mean value and the variance of R_n can be obtained from the formulæ derived in the preceding sections by substituting for $\bar{x}_N$, $\bar{y}_N$, $\mathbf{S}_x{}^2$, $\mathbf{S}_y{}^2$ and ρ in terms of N, p and q. Now it is easy to see that

$$\bar{y}_N = \frac{\sum_{i=1}^N y_i}{N} = \frac{N_1}{N} = p \tag{107}$$

$$\mathrm{S}_y^2 = \frac{\sum_{i=1}^{N} y_i^2 - N\bar{y}_N^2}{N-1}$$

$$= \frac{Np - Np^2}{N-1}$$

$$= \frac{N}{N-1} \cdot p \cdot q \qquad (108)$$

Similarly,

$$\bar{x}_N = q \qquad (109)$$

and

$$\mathrm{S}_x^2 = \frac{N}{N-1}\, pq \qquad (110)$$

Lastly,

$$\rho \mathrm{S}_x \mathrm{S}_y = \frac{\sum_{i=1}^{N} x_i y_i - N\bar{x}_N \bar{y}_N}{N-1}$$

$$= \frac{0 - Npq}{N-1} \qquad (111)$$

so that $\rho = -1$. On substituting in (12), we have

$$E\left(\frac{n_1}{n_2}\right) = \frac{N_1}{N_2}\left\{1 + \frac{N-n}{Nn}\left(\frac{N}{N-1} \cdot \frac{p}{q} + \frac{N}{N-1}\right)\right\}$$

$$= \frac{N_1}{N_2}\left\{1 + \frac{N-n}{N-1} \cdot \frac{1}{n}\left(\frac{p}{q} + 1\right)\right\}$$

$$= \frac{N_1}{N_2}\left\{1 + \frac{N-n}{N-1} \cdot \frac{1}{nq}\right\} \qquad (112)$$

It follows that when N is large, so that the finite multiplier can be taken to be unity, the relative bias in the ratio is given by $1/nq$.

The table below gives the values of n for different values of q in order that the relative bias in a ratio estimate may not exceed 2%.

q	..	·1	·2	·3	·4	·5
n	..	500	250	167	125	100

It is seen that n has ordinarily to be very large and particularly so when q is small, in order that the bias may be negligible.

There is one other point which needs to be emphasised. We have seen that when the relationship between y and x is a straight line passing through the origin, the bias vanishes. This is not so in the present case, for when y is 1, x is 0 and *vice versa*, and the regression line of y on x does not pass through the origin. This explains the need for a relatively larger value of n as compared to that in the case of quantitative characters in order that the ratio of the numbers in the two classes may give an unbiased estimate.

To obtain the first approximation to the variance of R_n, we substitute from (107), (108), (109), (110) and (111) in (24), giving

$$V_1\left(\frac{R_n}{R_N}\right) = \frac{N-n}{N-1}\cdot\frac{1}{n}\left(\frac{q}{p}+\frac{p}{q}+2\right)$$

$$= \frac{N-n}{N-1}\cdot\frac{1}{n}\left(\frac{1}{p}+\frac{1}{q}\right)$$

$$= \frac{N-n}{N-1}\cdot\frac{1}{npq} \tag{113}$$

4*a*.13 Extension to *k* Classes

The extension of the previous results when the population is divided into k classes is straightforward. Consider the ratio n_i/n_j, where n_i is the number in the sample in the i-th class and n_j in the j-th class, and assume that y takes the value 1 whenever it falls in class i and 0 elsewhere. Similarly, let x assume the value 1 when it falls in class j and 0 elsewhere. We then have

$$R_n = \frac{n_i}{n_j}, \qquad R_N = \frac{N_i}{N_j} \tag{114}$$

$$\bar{y}_N = p_i, \qquad \bar{x}_N = p_j \tag{115}$$

$$\mathbf{S}_y^2 = \frac{N}{N-1}\, p_i(1-p_i) \tag{116}$$

$$\mathbf{S}_x^2 = \frac{N}{N-1}\, p_j(1-p_j) \tag{117}$$

and

$$\rho S_y S_x = -\frac{N}{N-1} p_i p_j \tag{118}$$

so that

$$\rho = -\sqrt{\frac{p_i p_j}{(1-p_i)(1-p_j)}} \tag{119}$$

Substituting from the above in (12), we have

$$E\left(\frac{n_i}{n_j}\right) = \frac{N_i}{N_j}\left\{1 + \frac{N-n}{N}\cdot\frac{1}{n}\left(\frac{N}{N-1}\cdot\frac{p_j(1-p_j)}{p_j^2}\right.\right.$$

$$\left.\left. + \frac{N}{N-1}\cdot\frac{p_i p_j}{p_i p_j}\right)\right\}$$

$$= \frac{N_i}{N_j}\left\{1 + \frac{N-n}{N-1}\cdot\frac{1}{n}\left(\frac{1-p_j}{p_j} + 1\right)\right\}$$

$$= \frac{N_i}{N_j}\left\{1 + \frac{N-n}{N-1}\cdot\frac{1}{n}\cdot\frac{1}{p_j}\right\} \tag{120}$$

It will be seen that the formula is identical with (112) except that q is now replaced by p_j.

To obtain the variance, we substitute from (115) to (118) in (24) and obtain

$$V_1\left(\frac{R_n}{R_N}\right) = \frac{N-n}{N}\cdot\frac{1}{n}\left(\frac{N}{N-1}\cdot\frac{p_i(1-p_i)}{p_i^2}\right.$$

$$\left. + \frac{N}{N-1}\cdot\frac{p_j(1-p_j)}{p_j^2} + \frac{2N}{N-1}\cdot\frac{p_i p_j}{p_i p_j}\right)$$

$$= \frac{N-n}{N-1}\cdot\frac{1}{n}\left(\frac{1-p_i}{p_i} + \frac{1-p_j}{p_j} + 2\right)$$

$$= \frac{N-n}{N-1}\cdot\frac{1}{n}\left(\frac{1}{p_i} + \frac{1}{p_j}\right) \tag{121}$$

B. SAMPLING WITH VARYING PROBABILITIES OF SELECTION

4*b*.1 Ratio Estimate and its Variance

So far we have considered the theory of the ratio method of estimation for samples chosen by the method of simple random sampling. We shall now give the basic theory of the ratio estimate when sampling is carried out with replacement with varying probabilities of selection.

Let

$$z_i = \frac{y_i}{NP_i} = \bar{y}_N + \epsilon_i$$

$$v_i = \frac{x_i}{NP_i} = \bar{x}_N + \epsilon_i'$$

and

$$r_i = \frac{z_i}{v_i} = \frac{y_i}{x_i}$$

$$R_n = \frac{\bar{z}_n}{\bar{v}_n}$$

It is then easily shown that

$$E(z_i) = \bar{y}_N \qquad E(v_i) = \bar{x}_N$$

$$E(\bar{z}_n) = \bar{y}_N \qquad E(\bar{v}_n) = \bar{x}_N$$

$$E(\bar{\epsilon}_n^2) = \frac{\sigma_z^2}{n} = \frac{1}{n}\sum_{i=1}^{N} P_i (z_i - \bar{y}_N)^2$$

$$E(\bar{\epsilon}_n'^2) = \frac{\sigma_v^2}{n} = \frac{1}{n}\sum_{i=1}^{N} P_i (v_i - \bar{x}_N)^2$$

$$E(\bar{\epsilon}_n\bar{\epsilon}_n') = \frac{1}{n}\left\{\sum_{i=1}^{N} P_i (z_i - \bar{y}_N)\ (v_i - \bar{x}_N)\right\} = \frac{1}{n}\rho\sigma_z\sigma_v$$

On substituting in (9), it follows that

$$E_1(R_n) = R_N \left\{1 + \frac{1}{n}\left(\frac{\sigma_v^2}{\bar{x}_N^2} - \frac{\rho\sigma_z\sigma_v}{\bar{y}_N\bar{x}_N}\right)\right\} \tag{122}$$

Also, from (22), we have

$$V_1(R_n) = R_N^2 \frac{1}{n}\left(\frac{\sigma_z^2}{\bar{y}_N^2} + \frac{\sigma_v^2}{\bar{x}_N^2} - \frac{2\rho\sigma_z\sigma_v}{\bar{y}_N\bar{x}_N}\right) \tag{123}$$

This can be rewritten as

$$\begin{aligned} V_1(R_n) &= \frac{1}{n\bar{x}_N^2}(\sigma_z^2 + R_N^2\sigma_v^2 - 2R_N\rho\sigma_z\sigma_v) \\ &= \frac{1}{n\bar{x}_N^2}\left\{\sum_{i=1}^{N} P_i z_i^2 - \bar{y}_N^2 + R_N^2\left(\sum_{i=1}^{N} P_i v_i^2 - \bar{x}_N^2\right)\right. \\ &\qquad\qquad \left. - 2R_N\left(\sum_{i=1}^{N} P_i z_i v_i - \bar{y}_N\bar{x}_N\right)\right\} \\ &= \frac{1}{n\bar{x}_N^2}\left\{\sum_{i=1}^{N} P_i(z_i^2 + R_N^2 v_i^2 - 2R_N z_i v_i)\right\} \\ &= \frac{1}{n\bar{x}_N^2}\left\{\sum_{i=1}^{N} P_i(z_i - R_N v_i)^2\right\} \end{aligned} \tag{124}$$

or

$$V_1(R_n) = \frac{1}{nN^2\bar{x}_N^2}\left\{\sum_{i=1}^{N} \frac{1}{P_i}(y_i - R_N x_i)^2\right\} \tag{125}$$

It follows that

$$V_1(Y_R) = \frac{N^2}{n}\left\{\sum_{i=1}^{N} P_i(z_i - R_N v_i)^2\right\} \tag{126}$$

or

$$V_1(Y_R) = \frac{1}{n}\left\{\sum_{i=1}^{N} \frac{1}{P_i}(y_i - R_N x_i)^2\right\} \tag{127}$$

The estimates of the variances of R_n and Y_R are obtained directly from (40) and (41). We write

$$\text{Est. } V_1(R_n) = \frac{1}{n}\frac{1}{\bar{v}_n^2}\frac{1}{n-1}\sum^n (z_i - R_n v_i)^2 \qquad . \qquad (128)$$

and

$$\text{Est. } V_1(Y_R) = \frac{N^2}{n}\frac{1}{n-1}\sum^n (z_i - R_n v_i)^2 \qquad (129)$$

Finally, we note that when P_i is proportional to x_i so that $P_i = x_i/N\bar{x}_N$, we have

$$z_i = \frac{y_i}{NP_i} = \frac{y_i}{x_i}\bar{x}_N = r_i\bar{x}_N$$

$$v_i = \frac{x_i}{NP_i} = \bar{x}_N$$

$$\bar{z}_n = \bar{r}_n\bar{x}_N, \quad \bar{v}_n = \bar{x}_N$$

$$R_n = \bar{r}_n, \quad R_N = \bar{r}_N$$

and

$$Y_R = N\bar{z}_n = N\bar{x}_N \cdot \bar{r}_n$$

It follows that the sampling theory of the ratio estimate Y_R is identical with the sampling theory of the estimate $\bar{z}_n$. We therefore have from Section 2b.2,

$$E(Y_R) = NE(\bar{z}_n)$$
$$= N\bar{y}_N$$

showing that the ratio estimate for this case provides an unbiased estimate of the population total. Further, from (126), we have

$$V(Y_R) = \frac{N^2}{n} \cdot \bar{x}_N^2 \cdot \sum_{i=1}^{N} P_i(r_i - \bar{r}_N)^2$$

and, from (129),

$$\text{Est. } V(Y_R) = \frac{N^2\bar{x}_N^2}{n} \cdot \frac{1}{n-1}\sum^n (r_i - \bar{r}_n)^2$$

The method of forecasting acreage of principal crops in India provides a good example of the application of the theory presented in this section. Normally, area figures are collected by the village accountant, field by field, for all the villages within his jurisdiction. Such complete enumeration is, however, not available in time for making pre-harvest forecasts. These are consequently made on the basis of advance enumeration of a sample of villages selected with probability proportional to the cultivated area (including fallows) with which the area under a major crop is known to be highly correlated. The ratio method of estimation on the previous year's figures is used.

Example 4.4

Table 4.5 shows the total cultivated area during 1931 as also the area under wheat in two consecutive years 1936, 1937 for a sample of 34 villages in Lucknow subdivision (India). The villages were selected with replacement with probability proportional to the cultivated area (including fallows) as recorded in 1931. The total cultivated area in 1931 and the total area under wheat in 1936 for all the 170 villages in Lucknow subdivision were known to be 78019 and 21288 acres respectively. Estimate the area under wheat for the subdivision for the year 1937 using the ratio method of estimation and calculate the standard error of the estimate so made.

What would be the standard error of the estimate if the information for the previous year were not used?

TABLE 4.5

Values of Total Cultivated Area and of Area under Wheat in Two Consecutive Years for a Sample of 34 Villages in Lucknow Subdivision

Serial No. of Village	Total Cultivated Area in 1931 (Acres) 'a'	Area under Wheat		$l' = \frac{1000\,x}{a} = \frac{v}{0{\cdot}45894}$	$l = \frac{1000\,y}{a} = \frac{z}{0{\cdot}45894}$
		1936 (Acres) 'x'	1937 (Acres) 'y'		
(1)	(2)	(3)	(4)	(5)	(6)
1	401	75	52	187	130
2	634	163	149	257	235
3	1194	326	289	273	242
4	1770	442	381	250	215
5	1060	254	278	240	262
6	827	125	111	151	134
7	1737	559	634	322	365
8	1060	254	278	240	262
9	360	101	112	281	311
10	946	359	355	379	375
11	470	109	99	232	211
12	1625	481	498	296	306
13	827	125	111	151	134
14	96	5	6	52	63
15	1304	427	399	327	306
16	377	78	79	207	210
17	259	78	105	301	405
18	186	45	27	242	145
19	1767	564	515	319	291
20	604	238	249	394	412
21	701	92	85	131	121
22	524	247	221	471	422

TABLE 4.5—(*Contd.*)

Serial No. of Village	Total Cultivated Area in 1931 (Acres) 'a'	Area under Wheat		$l' = \frac{1000\,x}{a} = \frac{v}{0 \cdot 45894}$	$l = \frac{1000\,y}{a} = \frac{z}{0 \cdot 45894}$
		1936 (Acres) 'x'	1937 (Acres) 'y'		
(1)	(2)	(3)	(4)	(5)	(6)
23	571	134	133	235	233
24	962	131	144	136	150
25	407	129	103	317	253
26	715	192	179	269	250
27	845	663	330	785	391
28	1016	236	219	232	216
29	184	73	62	397	337
30	282	62	79	220	280
31	194	71	60	366	309
32	439	137	100	312	228
33	854	196	141	230	165
34	824	255	265	309	322
		Total	..	9511	8691
		Crude Sum of Squares	..	3166531	2505925
		Crude Sum of Products	..	2727616	

Let a_i denote the cultivated area in the i-th village and x_i and y_i the areas under wheat for the years 1936 and 1937 respectively. Then P_i the selection probability for the i-th village is given by $P_i = a_i/A$, where $A = \sum_{i=1}^{N} a_i = 78019$.

Also

$$z_i = \frac{y_i}{a_i} \cdot \frac{A}{N} \quad \text{and} \quad v_i = \frac{x_i}{a_i} \cdot \frac{A}{N}$$

For the sake of computational convenience $1000y_i/a_i$ and $1000x_i/a_i$ have been calculated instead of z_i and v_i. These are given in cols. 5 and 6 of Table 4.5 and denoted by l_i and l_i', where

$$l_i = \frac{1000N}{A} z_i = \frac{z_i}{0 \cdot 45894}$$

and

$$l_i' = \frac{1000N}{A} v_i = \frac{v_i}{0 \cdot 45894}$$

Now

$$R_n = \frac{\sum^n z_i}{\sum^n v_i} = \frac{\sum^n l_i}{\sum^n l_i'} = \frac{8691}{9511} = 0 \cdot 9138$$

Hence the estimate of the area under wheat in 1937 is given by

$$\begin{aligned} Y_R &= R_n X \\ &= (0 \cdot 9138)\ (21288) \\ &= 19453 \text{ acres} \end{aligned}$$

From Table 4.5,

$$\sum^n l_i^2 = 2505925, \ \sum^n l_i l_i' = 2727616 \text{ and } \sum^n l_i'^2 = 3166531$$

Hence

$$\begin{aligned} \sum^n (l_i - R_n l_i')^2 &= \sum^n l_i^2 - 2R_n \sum^n l_i l_i' + R_n^2 \sum^n l_i'^2 \\ &= 165082 \end{aligned}$$

or

$$\begin{aligned} \sum^n (z_i - R_n v_i)^2 &= \left(\frac{A}{1000N}\right)^2 \sum^n (l_i - R_n l_i')^2 \\ &= (0 \cdot 45894)^2\ (165082) \\ &= 34770 \end{aligned}$$

Hence

$$\begin{aligned} V(Y_R) &= \frac{N^2}{n} \cdot \frac{1}{n-1} \sum^n (z_i - R_n v_i)^2 \\ &= \frac{(170)^2}{(34)\ (33)} \cdot 34770 \\ &= 895600 \text{ acres}^2 \end{aligned}$$

$$\therefore \quad S.E. \quad Y_R = \sqrt{895600} = 946 \text{ acres}$$

If the information for the previous year had not been used. the estimate of the area under wheat in 1937 would have been

$$\hat{Y} = N\bar{z}_n$$

$$= (0\cdot45894) \cdot \frac{N}{n} \cdot \sum^{n} l_i$$

$$= \frac{(0\cdot45894)\,(170)\,(8691)}{34}$$

$$= 19943 \text{ acres}$$

and

$$V(\hat{Y}) = \frac{N^2}{n} \cdot \frac{1}{n-1} \cdot \sum^{n} (z_i - \bar{z}_n)^2$$

$$= \frac{N^2}{n} \cdot \frac{1}{n-1} \cdot \left(\frac{A}{1000N}\right)^2 \left\{\sum^{n} l_i^2 - \frac{\left(\sum^{n} l_i\right)^2}{n}\right\}$$

$$= \frac{(170)^2}{(34)\,(33)} \cdot (0\cdot45894)^2\,(2505925 - 2221573)$$

$$= \frac{850}{33}\,(0\cdot21063)\,(284352)$$

$$= 1542700 \text{ acres}^2$$

or

$$S.E.\quad \hat{Y} = 1242 \text{ acres}$$

The increase in efficiency in using the previous year's information

$$= \left(\frac{1542700}{895600} - 1\right)$$

$$= 72\cdot3\%$$

REFERENCES

1. David, F. N. and Neyman, J. (1938) "Extension of the Markoff Theorem on Least Squares," *Statist. Res. Mem.*, 2, 105–16.
2. Cochran, W. G. (1940) .. "The Estimation of the Yields of Cereal Experiments by Sampling for the Ratio of Grain to Total Produce," *Jour. Agr. Sci.*, **30**, 262–75.

3. Sukhatme, P. V. (1944) .. "Moments and Product Moments of Moment-Statistics for Samples of the Finite and Infinite Populations," *Sankhya*, **6**, 363–82.

4. Hansen, M. H., Hurwitz, W. N. and Gurney, M. (1946) "Problems and Methods of the Sample Survey of Business," *Jour. Amer. Statist. Assoc.*, **41**, 173–89.

5. Ayachit, G. R. (1953) .. "Some Aspects of Large-Scale Sample Surveys with Particular Reference to the Ratio Method of Estimation," *M.Sc. Thesis, Bombay University, Bombay.*

APPENDIX

Expected Values of Certain Higher Order Product Moments

We shall derive expressions for

$$\text{(i) } E(\bar{\epsilon}_n \bar{\epsilon}_n'^2), \quad \text{(ii) } E(\bar{\epsilon}_n \bar{\epsilon}_n'^3) \quad \text{and} \quad \text{(iii) } E(\bar{\epsilon}_n^2 \bar{\epsilon}_n'^2)$$

where

$$\epsilon_i = y_i - \bar{y}_N$$

$$\epsilon_i' = x_i - \bar{x}_N$$

and consequently

$$\bar{\epsilon}_n = \bar{y}_n - \bar{y}_N \quad \text{and} \quad \bar{\epsilon}_n' = \bar{x}_n - \bar{x}_N$$

Let

$$\sum_i^N \epsilon_i^a \epsilon_i'^\alpha = N\mu_{a,\alpha} \tag{1}$$

We may then write

$$\sum_{i \neq j}^N \epsilon_i^a \epsilon_j^b \epsilon_i'^\alpha \epsilon_j'^\beta = \sum_i^N \epsilon_i^a \epsilon_i'^\alpha \left(\sum_j^N \epsilon_j^b \epsilon_j'^\beta - \epsilon_i^b \epsilon_i'^\beta \right)$$

$$= N^2 \mu_{a,\alpha} \mu_{b,\beta} - N\mu_{a+b,\alpha+\beta} \tag{2}$$

Also

$$\sum_{i \neq j \neq k}^N \epsilon_i^a \epsilon_j^b \epsilon_k^c \epsilon_i'^\alpha \epsilon_j'^\beta \epsilon_k'^\gamma$$

$$\begin{aligned} &= \sum_{i \neq j}^N \epsilon_i^a \epsilon_j^b \epsilon_i'^\alpha \epsilon_j'^\beta \left(\sum_k^N \epsilon_k^c \epsilon_k'^\gamma - \epsilon_i^c \epsilon_i'^\gamma - \epsilon_j^c \epsilon_j'^\gamma \right) \\ &= N\mu_{c,\gamma} (N^2 \mu_{a,\alpha} \mu_{b,\beta} - N\mu_{a+b,\alpha+\beta}) \\ &\qquad - (N^2 \mu_{a+c,\alpha+\gamma} \mu_{b,\beta} - N\mu_{a+b+c,\alpha+\beta+\gamma}) \\ &\qquad - (N^2 \mu_{b+c,\beta+\gamma} \mu_{a,\alpha} - N\mu_{a+b+c,\alpha+\beta+\gamma}) \\ &= N^3 \mu_{a,\alpha} \mu_{b,\beta} \mu_{c,\gamma} - N^2 (\mu_{a+b,\alpha+\beta} \mu_{c,\gamma} \\ &\qquad + \mu_{a+c,\alpha+\gamma} \mu_{b,\beta} + \mu_{b+c,\beta+\gamma} \mu_{a,\alpha}) \\ &\qquad + 2N\mu_{a+b+c,\alpha+\beta+\gamma} \end{aligned} \tag{3}$$

Armed with these results and the theorem on expectations in Section 2*a*.9, the derivation of the expected values of higher order product moments becomes straightforward.

Thus

$$\begin{aligned} E(\bar{\epsilon}_n \bar{\epsilon}_n'^2) &= \frac{1}{n^3} E\left[\left(\sum_i^n \epsilon_i\right)\left(\sum_i^n \epsilon_i'\right)^2\right] \\ &= \frac{1}{n^3} E\left[\left(\sum_i^n \epsilon_i\right)\left(\sum_i^n \epsilon_i'^2 + \sum_{i \neq j}^n \epsilon_i' \epsilon_j'\right)\right] \\ &= \frac{1}{n^3} E\left[\sum_i^n \epsilon_i \epsilon_i'^2 + \sum_{i \neq j}^n \epsilon_i \epsilon_j'^2 + 2\sum_{i \neq j}^n \epsilon_i \epsilon_i' \epsilon_j' + \sum_{i \neq j \neq k}^n \epsilon_i \epsilon_j' \epsilon_k'\right] \qquad (4) \\ &= \frac{1}{n^3}\left[e_1 \sum_i^N \epsilon_i \epsilon_i'^2 + e_2 \sum_{i \neq j}^N (\epsilon_i \epsilon_j'^2 + 2\epsilon_i \epsilon_i' \epsilon_j') \right. \\ &\qquad \left. + e_3 \sum_{i \neq j \neq k}^N \epsilon_i \epsilon_j' \epsilon_k'\right] \\ &= \frac{1}{n^3}\left[e_1 N\mu_{12} + e_2(-N\mu_{12} - 2N\mu_{12}) + e_3(2N\mu_{12})\right] \\ &= \frac{N}{n^3}\left[e_1 - 3e_2 + 2e_3\right]\mu_{12} \\ &= \frac{(N-n)(N-2n)}{(N-1)(N-2)} \frac{\mu_{12}}{n^2} \qquad (5) \end{aligned}$$

It follows that

$$\begin{aligned} E(\bar{\epsilon}_n'^3) &= \frac{N}{n^3}(e_1 - 3e_2 + 2e_3)\,\mu_{03} \\ &= \frac{(N-n)(N-2n)}{(N-1)(N-2)} \frac{\mu_{03}}{n^2} \qquad (6) \end{aligned}$$

Similarly

$$E(\bar{\epsilon}_n \bar{\epsilon}_n'^3) = E\left[(\bar{\epsilon}_n')(\bar{\epsilon}_n \bar{\epsilon}_n'^2)\right]$$

Substituting from (4), we write

$$\begin{aligned} E(\bar{\epsilon}_n \bar{\epsilon}_n'^3) = \frac{1}{n^4} E\left[\left(\sum_i^n \epsilon_i'\right)\left(\sum_i^n \epsilon_i \epsilon_i'^2 + \sum_{i \neq j}^n \epsilon_i \epsilon_j'^2 + 2\sum_{i \neq j}^n \epsilon_i \epsilon_i' \epsilon_j' \right.\right. \\ \left.\left. + \sum_{i \neq j \neq k}^n \epsilon_i \epsilon_j' \epsilon_k'\right)\right] \end{aligned}$$

$$= \frac{1}{n^4}\, E\left[\sum_{i}^{n} \epsilon_i\epsilon_i'^3 + \sum_{i\neq j}^{n} (\epsilon_i\epsilon_j'^3 + 3\epsilon_i\epsilon_i'^2\epsilon_j' + 3\epsilon_i\epsilon_i'\epsilon_j'^2)\right.$$

$$\left. + \sum_{i\neq j\neq k}^{n} 3\,(\epsilon_i\epsilon_i'\epsilon_j'\epsilon_k' + \epsilon_i\epsilon_j'^2\epsilon_k') + \sum_{i\neq j\neq k\neq l}^{n} \epsilon_i\epsilon_j'\epsilon_k'\epsilon_l'\right]$$

$$= \frac{1}{n^4}\left[e_1 \sum_{i}^{N} \epsilon_i\epsilon_i'^3 + e_2 \sum_{i\neq j}^{N} (\epsilon_i\epsilon_j'^3 + 3\epsilon_i\epsilon_i'^2\epsilon_j' + 3\epsilon_i\epsilon_i'\epsilon_j'^2)\right.$$

$$\left. + e_3 \sum_{i\neq j\neq k}^{N} (3\epsilon_i\epsilon_i'\epsilon_j'\epsilon_k' + 3\epsilon_i\epsilon_j'^2\epsilon_k') + e_4 \sum_{i\neq j\neq k\neq l}^{N} \epsilon_i\epsilon_j'\epsilon_k'\epsilon_l'\right]$$

$$= \frac{1}{n^4}\left[e_1 \sum_{i}^{N} \epsilon_i\epsilon_i'^3 + e_2 \sum_{i\neq j}^{N} (\epsilon_i\epsilon_j'^3 + 3\epsilon_i\epsilon_i'^2\epsilon_j' + 3\epsilon_i\epsilon_i'\epsilon_j'^2)\right.$$

$$+ e_3 \sum_{i\neq j\neq k}^{N} (3\epsilon_i\epsilon_i'\epsilon_j'\epsilon_k' + 3\epsilon_i\epsilon_j'^2\epsilon_k')$$

$$\left. + e_4 \sum_{i\neq j\neq k}^{N} \epsilon_i\epsilon_j'\epsilon_k' \left(\sum_{l}^{N} \epsilon_l' - \epsilon_i' - \epsilon_j' - \epsilon_k'\right)\right]$$

$$= \frac{1}{n^4}\left[e_1 \sum_{i}^{N} \epsilon_i\epsilon_i'^3 + e_2 \sum_{i\neq j}^{N} (\epsilon_i\epsilon_j'^3 + 3\epsilon_i\epsilon_i'^2\epsilon_j' + 3\epsilon_i\epsilon_i'\epsilon_j'^2)\right.$$

$$+ e_3 \sum_{i\neq j\neq k}^{N} (3\epsilon_i\epsilon_i'\epsilon_j'\epsilon_k' + 3\epsilon_i\epsilon_j'^2\epsilon_k')$$

$$\left. - e_4 \sum_{i\neq j\neq k}^{N} (\epsilon_i\epsilon_i'\epsilon_j'\epsilon_k' + 2\epsilon_i\epsilon_j'^2\epsilon_k')\right]$$

Using the results (1), (2) and (3) above, we get

$$E(\bar{\epsilon}_n\bar{\epsilon}_n'^3) = \frac{1}{n^4}\left[e_1N\mu_{13} + e_2\{-N\mu_{13} - 3N\mu_{13} + 3\,(N^2\mu_{11}\mu_{02}\right.$$

$$\left. - N\mu_{13})\} + 3e_3\{4N\mu_{13} - 2N^2\mu_{11}\mu_{02}\} - e_4\{6N\mu_{13} - 3N^2\mu_{11}\mu_{02}\}\right]$$

$$= \frac{1}{n^4}\left[(e_1 - 7e_2 + 12e_3 - 6e_4)\,N\mu_{13}\right.$$

$$\left. + 3\,(e_2 - 2e_3 + e_4)\,N^2\mu_{11}\mu_{02}\right] \qquad (7)$$

$$= \frac{N-n}{n^3}\left[\frac{(N^2+N-6nN+6n^2)}{(N-1)(N-2)(N-3)}\,\mu_{13}\right.$$

$$\left. + \frac{3(n-1)(N-n-1)}{(N-1)(N-2)(N-3)}\,N\mu_{11}\mu_{02}\right] \quad (8)$$

It follows that

$$E(\bar{\epsilon}_n'^4) = \frac{1}{n^4}\Big[(e_1 - 7e_2 + 12e_3 - 6e_4)\,N\mu_{04}$$

$$+ 3(e_2 - 2e_3 + e_4)\,N^2\mu_{02}^2\Big] \quad (9)$$

or

$$= \frac{N-n}{n^3}\left[\frac{(N^2+N-6nN+6n^2)}{(N-1)(N-2)(N-3)}\,\mu_{04}\right.$$

$$\left. + \frac{3(n-1)(N-n-1)}{(N-1)(N-2)(N-3)}\,N\mu_{02}^2\right] \quad (10)$$

Lastly

$$E(\bar{\epsilon}_n^2\bar{\epsilon}_n'^2) = \frac{1}{n^4}\,E\Big[\Big(\sum_i^n \epsilon_i\Big)\Big\{\sum_i^n \epsilon_i\epsilon_i'^2 + \sum_{i\neq j}^n(\epsilon_i\epsilon_j'^2 + 2\,\epsilon_i\epsilon_i'\epsilon_j')$$

$$+ \sum_{i\neq j\neq k}^n \epsilon_i\epsilon_j'\epsilon_k'\Big\}\Big]$$

$$= \frac{1}{n^4}\,E\Big[\sum_i^n \epsilon_i^2\epsilon_i'^2 + \sum_{i\neq j}^n(2\epsilon_i\epsilon_j\epsilon_i'^2 + \epsilon_i^2\epsilon_j'^2 + 2\epsilon_i^2\epsilon_i'\epsilon_j'$$

$$+ 2\epsilon_i\epsilon_j\epsilon_i'\epsilon_j') + \sum_{i\neq j\neq k\neq l}^n \epsilon_i\epsilon_l\epsilon_j'\epsilon_k' + \sum_{i\neq j\neq k}^n$$

$$(\epsilon_i\epsilon_k\epsilon_j'^2 + 2\epsilon_i\epsilon_k\epsilon_i'\epsilon_j' + \epsilon_i^2\epsilon_j'\epsilon_k' + 2\epsilon_i\epsilon_j\epsilon_j'\epsilon_k')\Big]$$

$$= \frac{1}{n^4}\Big[e_1N\mu_{22} + e_2(N^2\mu_{20}\mu_{02} + 2N^2\mu_{11}^2 - 7N\mu_{22})$$

$$- e_3(2N^2\mu_{20}\mu_{02} + 4N^2\mu_{11}^2 - 12N\mu_{22})$$

$$- e_4\sum_{i\neq j\neq k}^N(\epsilon_i^2\epsilon_j'\epsilon_k' + 2\epsilon_i\epsilon_j\epsilon_j'\epsilon_k')\Big]$$

$$= \frac{1}{n^4}\Big[e_1N\mu_{22} + e_2(N^2\mu_{20}\mu_{02} + 2N^2\mu_{11}^2 - 7N\mu_{22})$$

$$- e_3(2N^2\mu_{20}\mu_{02} + 4N^2\mu_{11}^2 - 12N\mu_{22})$$

$$+ e_4(N^2\mu_{20}\mu_{02} + 2N^2\mu_{11}^2 - 6N\mu_{22})\Big]$$

$$= \frac{1}{n^4}\Big[(e_1 - 7e_2 + 12e_3 - 6e_4)\, N\mu_{22}$$

$$+ (e_2 - 2e_3 + e_4)\; N^2 (\mu_{20}\mu_{02} + 2\mu_{11}^2)\Big] \qquad (11)$$

$$= \frac{N-n}{n^3}\left[\frac{(N^2 + N - 6nN + 6n^2)}{(N-1)(N-2)(N-3)}\,\mu_{22}\right.$$

$$\left. + \frac{(n-1)(N-n-1)}{(N-1)(N-2)(N-3)}\, N(\mu_{20}\mu_{02} + 2\mu_{11}^2)\right] \qquad (12)$$

CHAPTER V

REGRESSION METHOD OF ESTIMATION

5.1 Simple Regression

In this chapter, we shall consider the regression method of estimating the population total (or mean) of the character y under study. Suppose, as previously, that the population is divided into k classes with, say, N_i units having the value x_i each $(i = 1, 2, \ldots, k)$, and that a simple random sample of n is drawn from the population N. Suppose, further, that in repeated samples of n, the number of units having the value x_i is fixed, say given by n_i $(i = 1, 2, \ldots, k)$. In simple regression we postulate a procedure of sampling in which $n_1, n_2, \ldots, n_k$ units are drawn from their respective classes with replacement. Further, we assume that the mean value of y for a given x is linear in x and $V(y \mid x)$ is constant. In other words, y is of the form

$$y_{ij} = \alpha + \beta x_i + \epsilon_{ij} \tag{1}$$

where

$$E(\epsilon_{ij} \mid i) = 0$$

and

$$E(\epsilon_{ij}^2 \mid i) = \text{constant, say, } \gamma \tag{2}$$

Summing up (1) over all the N units in the population, we have

$$Y = N\alpha + \beta \sum_{i=1}^{k} N_i x_i \tag{3}$$

When α and β are estimated from the sample and the population total of x is known, the right-hand side of (3) provides an estimate of the total. This estimate is known as the *simple regression estimate*. To distinguish it from the estimate of the population total based on the ratio and the simple mean methods, it will be denoted by Y_l, and the mean by $\bar{y}_l$.

5.2 Simple Regression Estimate and its Variance

We have seen in Section $2a.3$ that the best unbiased linear estimate and its variance are given by the Markoff method of estimation. Suppose that Y_l is given by

$$Y_l = \sum_{i=1}^{k} n_i \lambda_i \bar{y}_{n_i} \tag{4}$$

where λ_i's are to be chosen so as to satisfy the conditions of the Markoff Theorem. The condition that Y_l should be an unbiased estimate of Y gives

$$E\left\{\sum_{i=1}^{k} n_i \lambda_i \bar{y}_{n_i}\right\} = \sum_{i=1}^{k} N_i \bar{y}_{N_i}$$

On substituting from (1), we obtain

$$\sum_{i=1}^{k} n_i \lambda_i (\alpha + \beta x_i) = \sum_{i=1}^{k} N_i (\alpha + \beta x_i)$$

which can be written as

$$\sum_{i=1}^{k} (n_i \lambda_i - N_i)(\alpha + \beta x_i) = 0 \tag{5}$$

The second condition of the Markoff method of estimation is that the variance of the estimate should be minimum. Now the variance of Y_l when $n_1, n_2, \ldots, n_k$ are fixed and sampling is carried out with replacement with constant variance γ in each class, is given by

$$V(Y_l \mid n_1, n_2, \ldots, n_k) = \sum_{i=1}^{k} n_i^2 \lambda_i^2 \frac{\sigma_i^2}{n_i}$$

where

$$\sigma_i^2 = \frac{\sum_{j=1}^{N_i} (y_{ij} - \bar{y}_{N_i})^2}{N_i} = \gamma \tag{6}$$

Hence

$$V(Y_l \mid n_1, n_2, \ldots, n_k) = \gamma \sum_{i=1}^{k} n_i \lambda_i^2 \tag{7}$$

To minimize (7) subject to the condition (5), we shall use the Lagrangian method of multipliers and form a function ϕ given by

$$\phi = \gamma \sum_{i=1}^{k} n_i\lambda_i^2 - \mu \sum_{i=1}^{k} (n_i\lambda_i - N_i)(\alpha + \beta x_i) \tag{8}$$

where μ is the Lagrangian constant.

On differentiating ϕ with respect to λ_i, α and β and equating to zero, we obtain

$$\frac{\partial\phi}{\partial\lambda_i} \equiv 2n_i\lambda_i\gamma - \mu n_i(\alpha + \beta x_i) = 0 \qquad (i = 1, 2, \ldots, k) \tag{9}$$

$$\frac{\partial\phi}{\partial\alpha} \equiv -\mu \sum_{i=1}^{k} (n_i\lambda_i - N_i) = 0 \tag{10}$$

and

$$\frac{\partial\phi}{\partial\beta} \equiv -\mu \sum_{i=1}^{k} x_i(n_i\lambda_i - N_i) = 0 \tag{11}$$

From (9) we have

$$\lambda_i = \frac{\mu(\alpha + \beta x_i)}{2\gamma} = \frac{\alpha' + \beta' x_i}{\gamma} \tag{12}$$

where

$$\alpha' = \frac{\mu\alpha}{2} \text{ and } \beta' = \frac{\mu\beta}{2} \tag{13}$$

Substituting from (12) for λ_i's in (10) and (11), we obtain

$$\frac{n}{\gamma}(\alpha' + \beta'\bar{x}_n) = N \tag{14}$$

and

$$\frac{n}{\gamma}\alpha'\bar{x}_n + \frac{\beta'}{\gamma}\sum_{i=1}^{k} n_i x_i^2 = N\bar{x}_N \tag{15}$$

Solving for α' and β', we get

$$\beta' = \frac{N\gamma\,(\bar{x}_N - \bar{x}_n)}{\sum_{i=1}^{k} n_i\,(x_i - \bar{x}_n)^2} \tag{16}$$

and

$$\alpha' = N\gamma\left\{\frac{1}{n} - \frac{\bar{x}_n\,(\bar{x}_N - \bar{x}_n)}{\sum_{i=1}^{k} n_i\,(x_i - \bar{x}_n)^2}\right\} \tag{17}$$

On substituting for β' and α' from (16) and (17) in (12), we obtain

$$\lambda_i = N\left\{\frac{1}{n} + \frac{(\bar{x}_N - \bar{x}_n)}{\sum_{i=1}^{k} n_i\,(x_i - \bar{x}_n)^2}\,(x_i - \bar{x}_n)\right\} (i=1, 2, \ldots, k) \tag{18}$$

Hence, from (4),

$$Y_l = N\left\{\bar{y}_n + \frac{\sum_{i=1}^{k} n_i\bar{y}_{n_i}\,(x_i - \bar{x}_n)}{\sum_{i=1}^{k} n_i\,(x_i - \bar{x}_n)^2}\,(\bar{x}_N - \bar{x}_n)\right\} \tag{19}$$

Estimating Y from (3) by

$$N\hat{\alpha} + N\bar{x}_N\hat{\beta}$$

and equating with the right-hand side of (19), we have

$$\hat{\alpha} = \bar{y}_n - \hat{\beta}\bar{x}_n \tag{20}$$

and

$$\hat{\beta} = \frac{\sum_{i=1}^{k} n_i\bar{y}_{n_i}\,(x_i - \bar{x}_n)}{\sum_{i=1}^{k} n_i\,(x_i - \bar{x}_n)^2} \tag{21}$$

To obtain the variance of Y_l, we substitute for λ_i from (18) in (7). We then have

$$V(Y_l) = \gamma \sum_{i=1}^{k} n_i\lambda_i^2$$

$$= N^2\gamma \sum_{i=1}^{k} n_i \left\{ \frac{1}{n^2} + \frac{2}{n} \cdot \frac{(x_i - \bar{x}_n)}{\sum\limits_{i=1}^{k} n_i\, (x_i - \bar{x}_n)^2}\, (\bar{x}_N - \bar{x}_n) + \frac{(x_i - \bar{x}_n)^2}{\left[\sum\limits_{i=1}^{k} n_i\, (x_i - \bar{x}_n)^2\right]^2}\, (\bar{x}_N - \bar{x}_n)^2 \right\}$$

Clearly, the middle term is zero, giving us

$$V(Y_t) = N^2\gamma \left\{ \frac{1}{n} + \frac{(\bar{x}_N - \bar{x}_n)^2}{\sum\limits_{i=1}^{k} n_i\, (x_i - \bar{x}_n)^2} \right\} \tag{22}$$

which can also be written as

$$V(Y_t) = \frac{N^2\gamma}{n} \left\{ 1 + \frac{(m_1 - \mu_1)^2}{m_2} \right\} \tag{23}$$

where m_1 is the sample mean $\bar{x}_n$, m_2 denotes the second moment of the sample about its mean and μ_1 the population mean of x.

Pooling (6) over k classes gives

$$\gamma = \frac{\sum\limits_{i=1}^{k} \sum\limits_{j=1}^{N_i} (y_{ij} - a - \beta x_i)^2}{\sum\limits_{i=1}^{k} N_i} \tag{24}$$

From (20) and (21), we have the identities:

$$a \equiv \bar{y}_N - \beta \bar{x}_N \tag{25}$$

and

$$\beta \equiv \frac{\sum\limits_{i=1}^{k} \sum\limits_{j=1}^{N_i} y_{ij}\, (x_i - \bar{x}_N)}{\sum\limits_{i=1}^{k} N_i\, (x_i - \bar{x}_N)^2}$$

or

$$= \frac{\sum\limits^{N} y\, (x - \bar{x}_N)}{\sum\limits^{N} (x - \bar{x}_N)^2}$$

where the summations in both numerator and denominator are carried out over all the N pairs (x_i, y_{ij}),

$$= \rho \frac{\sigma_y}{\sigma_x}, \text{ say} \tag{26}$$

On substituting for α and β from (25) and (26) in (24), we get

$$\gamma = \frac{1}{N} \sum^{N} \left\{ y - \bar{y}_N - \rho \frac{\sigma_y}{\sigma_x} (x - \bar{x}_N) \right\}^2$$

$$= \frac{1}{N} \left[\sum^{N} (y - \bar{y}_N)^2 - \rho^2 \frac{\sigma_y^2}{\sigma_x^2} \sum^{N} (x - \bar{x}_N)^2 \right]$$

$$= \sigma_y^2 (1 - \rho^2) \tag{27}$$

We may, therefore, write (22) as

$$V(Y_l \mid n_1, n_2, \ldots, n_k) = N^2 \sigma_y^2 (1 - \rho^2) \times \left\{ \frac{1}{n} + \frac{(\bar{x}_N - \bar{x}_n)^2}{\sum_{i=1}^{k} n_i (x_i - \bar{x}_n)^2} \right\} \tag{28}$$

It will be noticed that the variance consists of two terms. The first term represents N^2 times the residual variance of the mean of a simple random sample of n, estimated from the regression line, when β is known; while the second represents the increase in the variance of the estimate when β is determined from the sample. As will be shown in Section 5.4, the latter contribution is of an order $1/n^2$, so that if n is large, the variance of the regression estimate may be regarded as being given by the first term only. It should be pointed out, however, that the sampling is assumed to be carried out with replacement. For a simple random sample drawn without replacement, exact results are not available but it is surmised that the effect will be approximately to multiply the above expression by the finite multiplier $(N - n)/N$.

5.3 Estimation of the Variance of Simple Regression Estimate

To evaluate (22) we require the estimate of γ. A straightforward method of finding this is to substitute for α and β their sample estimates in (24) and calculate the expectation. Consider then a quantity **Q** given by

$$\mathbf{Q} = \sum_{i=1}^{k} \sum_{j}^{n_i} \{y_{ij} - \bar{y}_n - \hat{\beta}(x_i - \bar{x}_n)\}^2 \tag{29}$$

and calculate its conditional expectation for fixed $n_1, n_2, \ldots, n_k$. This is best done by expressing **Q** as a function of the ϵ's, which are defined by (1).

From (1) we have

$$\bar{y}_{n_i} = \alpha + \beta x_i + \bar{\epsilon}_{n_i} \tag{30}$$

whence

$$\sum_{i=1}^{k} n_i \bar{y}_{n_i} = n\alpha + \beta n \bar{x}_n + \sum_{i=1}^{k} n_i \bar{\epsilon}_{n_i}$$

i.e.,

$$\bar{y}_n = \alpha + \beta \bar{x}_n + \frac{\sum_{i=1}^{k} n_i \bar{\epsilon}_{n_i}}{n} \tag{31}$$

Also

$$\hat{\beta} = \frac{\sum_{i=1}^{k} n_i \bar{y}_{n_i} (x_i - \bar{x}_n)}{\sum_{i=1}^{k} n_i (x_i - \bar{x}_n)^2}$$

$$= \frac{\sum_{i=1}^{k} n_i (\alpha + \beta x_i + \bar{\epsilon}_{n_i})(x_i - \bar{x}_n)}{\sum_{i=1}^{k} n_i (x_i - \bar{x}_n)^2}$$

$$= \frac{\beta \sum_{i=1}^{k} n_i x_i (x_i - \bar{x}_n)}{\sum_{i=1}^{k} n_i (x_i - \bar{x}_n)^2} + \frac{\sum_{i=1}^{k} n_i \bar{\epsilon}_{n_i} (x_i - \bar{x}_n)}{\sum_{i=1}^{k} n_i (x_i - \bar{x}_n)^2}$$

$$= \beta + \frac{\sum_{i=1}^{k} n_i \bar{\epsilon}_{n_i} (x_i - \bar{x}_n)}{\sum_{i=1}^{k} n_i (x_i - \bar{x}_n)^2} \tag{32}$$

On substituting in (29) from (1), (31) and (32), we get

$$Q = \sum_{i=1}^{k}\sum_{j}^{n_i}\left\{\left(\epsilon_{ij} - \frac{\sum_{r=1}^{k} n_r\bar{\epsilon}_{n_r}}{n}\right) - \frac{(x_i - \bar{x}_n)\sum_{s=1}^{k} n_s(x_s - \bar{x}_n)\bar{\epsilon}_{n_s}}{\sum_{i=1}^{k} n_i(x_i - \bar{x}_n)^2}\right\}^2 \tag{33}$$

Setting $\sum_{i=1}^{k} n_i(x_i - \bar{x}_n)^2 = nm_2$ and expanding the right-hand side, we obtain

$$\begin{aligned}
Q = {} & \sum_{i=1}^{k}\sum_{j}^{n_i}\left(\epsilon_{ij} - \frac{\sum_{r=1}^{k} n_r\bar{\epsilon}_{n_r}}{n}\right)^2 \\
& + \sum_{i=1}^{k}\sum_{j}^{n_i}\frac{(x_i - \bar{x}_n)^2\left\{\sum_{s=1}^{k} n_s(x_s - \bar{x}_n)\bar{\epsilon}_{n_s}\right\}^2}{n^2m_2^2} \\
& - 2\sum_{i=1}^{k}\sum_{j}^{n_i}\left(\epsilon_{ij} - \frac{\sum_{r=1}^{k} n_r\bar{\epsilon}_{n_r}}{n}\right) \\
& \times \frac{(x_i - \bar{x}_n)\sum_{s=1}^{k} n_s(x_s - \bar{x}_n)\bar{\epsilon}_{n_s}}{nm_2} \\
= {} & \sum_{i=1}^{k}\sum_{j}^{n_i}\epsilon_{ij}^2 - \frac{1}{n}\left(\sum_{r=1}^{k} n_r\bar{\epsilon}_{n_r}\right)^2 - \frac{\left\{\sum_{s=1}^{k} n_s\bar{\epsilon}_{n_s}(x_s - \bar{x}_n)\right\}^2}{nm_2} \\
= {} & \sum_{i=1}^{k}\sum_{j}^{n_i}\epsilon_{ij}^2 - \frac{1}{n}\left\{\sum_{i=1}^{k} n_i^2\bar{\epsilon}_{n_i}^2 + \sum_{i\neq l=1}^{k} n_in_l\bar{\epsilon}_{n_i}\bar{\epsilon}_{nl}\right\} \\
& - \frac{1}{nm_2}\left\{\sum_{i=1}^{k} n_i^2\bar{\epsilon}_{n_i}^2(x_i - \bar{x}_n)^2\right. \\
& \left. + \sum_{i\neq l=1}^{k} n_in_l\bar{\epsilon}_{n_i}\bar{\epsilon}_{nl}(x_i - \bar{x}_n)(x_l - \bar{x}_n)\right\}
\end{aligned} \tag{34}$$

Taking expectations of both sides in (34) and noting that

$$E(\epsilon_{ij}^2) = \gamma$$

$$E(\bar{\epsilon}_{n_i}^2) = \frac{\gamma}{n_i}$$

and

$$E(\bar{\epsilon}_{n_i}\bar{\epsilon}_{n_l}) = 0$$

we obtain

$$E(\mathbf{Q}) = (n-2)\gamma \tag{35}$$

It follows that

$$\text{Est. } \gamma = \frac{\mathbf{Q}}{(n-2)} \tag{36}$$

On substituting in (22), we obtain

$$\text{Est. } V(Y_I \mid n_1, n_2, \ldots, n_k) = \frac{N^2\mathbf{Q}}{n-2} \times \left\{\frac{1}{n} + \frac{(\bar{x}_N - \bar{x}_n)^2}{\sum\limits_{i=1}^{k} n_i (x_i - \bar{x}_n)^2}\right\} \tag{37}$$

5.4 Expected Value of the Sampling Variance of Simple Regression Estimate

The expression for the sampling variance given in (23) or (28) depends upon the x's that have turned up in the sample, and so cannot be used for purposes like comparing the precision of the regression estimate with other estimates. We therefore proceed to obtain the expected value of the variance over all samples of size n.

Since the first term in (23) is independent of the x's, the problem reduces to that of determining the expected value of

$$\frac{(m_1 - \mu_1)^2}{m_2}$$

Without loss of generality we may assume μ_1 to be zero. Obviously, the expected value of m_1^2/m_2 can be determined only approximately, since both the numerator and the denominator are random variables. Following Section 4*a*.3, we may write

$$E\left(\frac{m_1^2}{m_2}\right) = E\left\{\frac{m_1^2}{\mu_2}\left(1 - \frac{m_2 - \mu_2}{\mu_2} + \frac{(m_2 - \mu_2)^2}{\mu_2^2} - \ldots\right)\right\} \tag{38}$$

where μ_2 denotes the second moment of x in the population. The expected value of the terms in $m_1^2\,(m_2 - \mu_2)^3/\mu_2^4$ and higher powers will be of an order smaller than $1/n$, and can be neglected if n is assumed to be reasonably large. We, therefore, write to terms of order $1/n^2$

$$E\left(\frac{m_1^2}{m_2}\right) = E\left(\frac{m_1^2}{\mu_2}\right) - E\left\{\frac{m_1^2\,(m_2 - \mu_2)}{\mu_2^2}\right\} + E\left\{\frac{m_1^2\,(m_2 - \mu_2)^2}{\mu_2^3}\right\} \tag{39}$$

The values of the expressions on the right-hand side have been tabulated for ready application (Sukhatme, 1944). A reference to these formulæ gives

$$E\,(m_1^2 m_2) = N\mu_4\left[\frac{(e_1 - 3e_2 + 2e_3)}{n^3} - \frac{(e_1 - 7e_2 + 12e_3 - 6e_4)}{n^4}\right] + N^2\mu_2^2\left[\frac{(e_2 - e_3)}{n^3} - 3\,\frac{(e_2 - 2e_3 + e_4)}{n^4}\right]$$

where

$$e_j = \frac{n\,(n-1)\,(n-2)\ldots(n-j+1)}{N\,(N-1)\,(N-2)\ldots(N-j+1)}$$

Neglecting terms of order smaller than $1/n^2$ and retaining terms in $1/N$ and $1/N^2$ for completeness, we obtain

$$E\,(m_1^2 m_2) = \mu_4\left(\frac{1}{n^2} - \frac{3}{nN} + \frac{2}{N^2}\right) + \mu_2^2\left(\frac{1}{n} - \frac{4}{n^2} - \frac{1}{N} + \frac{10}{nN} - \frac{6}{N^2}\right) \tag{40}$$

Similarly, on neglecting terms involving $1/n^3$ and higher powers of $1/n$ in the formula for $E\,(m_1^2 m_2^2)$ as tabulated by the author (1944), we have

$$E\,(m_1^2 m_2^2) = \mu_4\mu_2\left(\frac{3}{n^2} - \frac{8}{nN} + \frac{5}{N^2}\right) + \mu_3^2\left(\frac{2}{n^2} - \frac{4}{nN} + \frac{2}{N^2}\right)$$

$$+ \mu_2^3\left(\frac{1}{n} - \frac{9}{n^2} + \frac{21}{nN} - \frac{1}{N} - \frac{12}{N^2}\right) \quad (41)$$

Substituting in (39) from (40) and (41), and using the known result that

$$E\,(m_1^2) = \mu_2\left(\frac{1}{n} - \frac{1}{N}\right)$$

we have

$$E\left(\frac{m_1^2}{m_2}\right) \doteq \left(\frac{1}{n} - \frac{1}{N}\right)$$

$$\times \left\{1 + \frac{3}{n} - \frac{6}{N} + \frac{\beta_2}{N} + 2\beta_1\left(\frac{1}{n} - \frac{1}{N}\right)\right\} \quad (42)$$

where

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} \text{ and } \beta_2 = \frac{\mu_4}{\mu_2^2}$$

If the x's can be assumed to be normally distributed, so that $\beta_1 = 0$ and $\beta_2 = 3$, we can write the expression for $E\,(m_1^2/m_2)$ in the exact form as

$$E\left(\frac{m_1^2}{m_2}\right) = \frac{1}{n}\left(1 - \frac{3}{n}\right)^{-1} = \frac{1}{n-3} \quad (43)$$

The variance of Y_l to terms in $1/n^2$ is thus approximated by

$$V\,(Y_l) \cong N^2\left[\frac{\sigma_y^2\,(1-\rho^2)}{n}\left\{1 + \frac{1}{n} - \frac{1}{N}\right\}\right] \quad (44)$$

and to terms in $1/n$ by simply

$$V\,(Y_l) \cong \frac{N^2\sigma_y^2\,(1-\rho^2)}{n} \quad (45)$$

To obtain the variance of the mean $\bar{y}_l$, we have only to divide the expression on the right-hand side of (44) or (45) by N^2.

5.5 Weighted Regression

In developing the preceding theory for simple regression, we have assumed that: (*a*) the regression is linear, (*b*) the deviation from the regression line has a constant variance, and (*c*) sampling within a class is carried out with replacement. These assumptions may not hold in actual practice. In this section, we shall extend the theory to the case where the relationship is linear, the deviation from the regression in any class has a known variance and sampling within a class is carried out without replacement. The extension is due to Hasel (1942) and Cochran (1942).

Let Y_{wl} denote the weighted linear regression estimate of Y and suppose that it is represented by the following linear function of observations:

$$Y_{wl} = \sum_{i=1}^{k} n_i\lambda_i\bar{y}_{n_i} \tag{46}$$

where λ_i's ($i = 1, 2, \ldots, k$) are to be so determined that: (i) Y_{wl} is an unbiased estimate of Y, and (ii) its variance is minimum. Now, the first condition gives

$$E(Y_{wl}) = Y$$

i.e.,

$$\sum_{i=1}^{k} n_i\lambda_i(\alpha + \beta x_i) = \sum_{i=1}^{k} N_i(\alpha + \beta x_i)$$

or

$$\sum_{i=1}^{k} (n_i\lambda_i - N_i)(\alpha + \beta x_i) = 0 \tag{47}$$

The variance of Y_{wl} for fixed x's is clearly given by

$$V(Y_{wl} \mid n_1, n_2, \ldots, n_k) = \sum_{i=1}^{k} n_i^2\lambda_i^2 \frac{N_i - n_i}{N_i} \cdot \frac{S_i^2}{n_i}$$

$$= \sum_{i=1}^{k} \frac{n_i^2\lambda_i^2}{w_i} \tag{48}$$

where

$$w_i = \frac{n_i N_i}{\mathbf{S}_i^2 (N_i - n_i)} \tag{49}$$

and

$$\mathbf{S}_i^2 = \frac{\sum_{j=1}^{N_i} (y_{ij} - \bar{y}_{N_i})^2}{N_i - 1}$$

To minimize (48) subject to condition (47), we form the function ϕ given by

$$\phi = \sum_{i=1}^{k} \frac{n_i^2 \lambda_i^2}{w_i} - \mu \sum_{i=1}^{k} (n_i \lambda_i - N_i)\, (\alpha + \beta x_i) \tag{50}$$

On differentiating ϕ with respect to λ_i, α and β and equating to zero, we obtain

$$\frac{\partial \phi}{\partial \lambda_i} \equiv \frac{2 n_i^2 \lambda_i}{w_i} - \mu n_i (\alpha + \beta x_i) = 0 \qquad (i = 1, 2, \ldots, k) \tag{51}$$

whence

$$\lambda_i = \mu (\alpha + \beta x_i) \cdot \frac{w_i}{2 n_i} \tag{52}$$

$$\frac{\partial \phi}{\partial \alpha} \equiv -\mu \sum_{i=1}^{k} (n_i \lambda_i - N_i) = 0 \tag{53}$$

and

$$\frac{\partial \phi}{\partial \beta} \equiv -\mu \sum_{i=1}^{k} x_i (n_i \lambda_i - N_i) = 0 \tag{54}$$

Substituting for λ_i's from (52) in (53) and (54), we obtain

$$\alpha' W + \beta' W \bar{x}_w = N \tag{55}$$

and

$$\alpha' W \bar{x}_w + \beta' \sum_{i=1}^{k} w_i x_i^2 = N \bar{x}_N \tag{56}$$

where

$$\alpha' = \frac{\mu \alpha}{2}, \quad \beta' = \frac{\mu \beta}{2} \tag{57}$$

and

$$W = \sum_{i=1}^{k} w_i, \quad \bar{x}_w = \frac{\sum_{i=1}^{k} w_i x_i}{W} \tag{58}$$

Solving (55) and (56) for α' and β', we get

$$\hat{\beta}' = \frac{N(\bar{x}_N - \bar{x}_w)}{\sum_{i=1}^{k} w_i x_i^2 - W\bar{x}_w^2} = \frac{N(\bar{x}_N - \bar{x}_w)}{W s_{wx}^2} \tag{59}$$

and

$$\hat{a}' = \frac{N}{W} - \frac{N\bar{x}_w(\bar{x}_N - \bar{x}_w)}{\sum_{i=1}^{k} w_i x_i^2 - W\bar{x}_w^2} = \frac{N}{W} - \frac{N\bar{x}_w(\bar{x}_N - \bar{x}_w)}{W s_{wx}^2} \tag{60}$$

where

$$s_{wx}^2 = \frac{\sum_{i=1}^{k} w_i x_i^2 - W\bar{x}_w^2}{W} \tag{61}$$

On substituting for α' and β' in (52), we obtain

$$\lambda_i = \frac{Nw_i}{Wn_i}\left\{1 + \frac{\bar{x}_N - \bar{x}_w}{s_{wx}^2}(x_i - \bar{x}_w)\right\} \tag{62}$$

Hence, from (46), we have

$$Y_{wl} = \frac{N}{W}\sum_{i=1}^{k} w_i \bar{y}_{n_i}\left\{1 + \frac{\bar{x}_N - \bar{x}_w}{s_{wx}^2}(x_i - \bar{x}_w)\right\}$$

which may be written as

$$Y_{wl} = N[\bar{y}_w + \hat{\beta}(\bar{x}_N - \bar{x}_w)] \tag{63}$$

where

$$\bar{y}_w = \frac{1}{W}\sum_{i=1}^{k} w_i \bar{y}_{n_i} \tag{64}$$

and

$$\hat{\beta} = \frac{1}{W s_{wx}^2}\left\{\sum_{i=1}^{k} w_i x_i \bar{y}_{n_i} - W\bar{x}_w \bar{y}_w\right\} \tag{65}$$

To obtain the variance of Y_{wl}, we substitute for λ_i's from (62) in (48) and obtain

$$V(Y_{wl}) = \frac{N^2}{W}\left[1 + \frac{(\bar{x}_N - \bar{x}_w)^2}{s_{wx}^2}\right] \tag{66}$$

It can be verified that when sampling within a class is carried out with replacement and $V(y \mid x)$ is a constant, formulæ (63) and (66) reduce to those appropriate for the simple regression estimate. For, we have in this case

$$w_i = \frac{n_i}{\gamma}, \quad W = \frac{n}{\gamma}$$

$$\bar{x}_w = \bar{x}_n, \quad s_{wx}^2 = \frac{1}{n}\sum_{i=1}^{k} n_i (x_i - \bar{x}_n)^2$$

and

$$\hat{\beta} = \frac{\sum\limits_{i=1}^{k} n_i \bar{y}_{n_i} (x_i - \bar{x}_n)}{\sum\limits_{i=1}^{k} n_i (x_i - \bar{x}_n)^2}$$

and on substitution we find that (63) and (66) reduce to (19) and (22) respectively.

An examination of the expression for Y_{wl} in (63) shows that a knowledge of the true weights w_i is not necessary for calculating Y_{wl}; numbers proportional to w_i are sufficient for the purpose. The variance of Y_{wl}, on the other hand, requires a knowledge of the true weights w_i. This raises a practical difficulty, since the true weights w_i are rarely known. Often, however, the relationship between the variance of y for a given x, and x can be guessed, and numbers proportional to w_i can be known. It is, therefore, important to investigate the form which the variance of Y_{wl} takes for certain well-known relationships between $V(y \mid x)$ and x. We shall consider only the simplest situation where $V(y \mid x)$ is proportional to x.

Let

$$V(y \mid x) = \gamma x \tag{67}$$

It follows then from (49) that

$$\begin{aligned} w_i &= \frac{n_i}{V(y \mid x)} \cdot \frac{N_i - 1}{N_i - n_i} \\ &= \frac{n_i}{\gamma x_i} \cdot \frac{N_i - 1}{N_i - n_i} \\ &= \frac{w_i'}{\gamma} \end{aligned} \tag{68}$$

where

$$w_i' = \frac{n_i}{x_i} \cdot \frac{N_i - 1}{N_i - n_i} \tag{69}$$

and

$$W = \frac{1}{\gamma} \sum_{i=1}^{k} w_i' = \frac{W'}{\gamma} \tag{70}$$

On substituting in (66), we have

$$V(Y_{wl}) = N^2\gamma \left[\frac{1}{W'} + \frac{(\bar{x}_N - \bar{x}_w)^2}{\sum_{i=1}^{k} w_i' (x_i - \bar{x}_w)^2} \right] \tag{71}$$

which now depends only on the known numbers w_i' and the constant γ.

5.6 Estimation of the Variance of Weighted Regression Estimate

We shall consider this problem for the case where the variance of y for a given x is proportional to x. Since the variance of the weighted regression estimate for this case is given by (71), the problem reduces to that of the estimation of γ.

As in Section 5.3, we shall start with the quantity **Q** defined by

$$\mathbf{Q} = \sum_{i=1}^{k} \sum_{j}^{n_i} \frac{w_i'}{n_i} \left[y_{ij} - \bar{y}_w - \hat{\beta} (x_i - \bar{x}_w) \right]^2 \tag{72}$$

and proceed to calculate its expectation for a given set of x's. A straightforward method of doing this is to express (72) in terms of ϵ_{ij} and x's, and then take expectations. Now ϵ_{ij} is defined by (1), namely,

$$y_{ij} = \alpha + \beta x_i + \epsilon_{ij}$$

where

$$E(\epsilon_{ij} \mid i) = 0$$

and

$$\begin{aligned} E(\epsilon_{ij}^2 \mid i) &= V(y \mid x) \\ &= \gamma x_i \end{aligned} \tag{73}$$

From (1), we have

$$\bar{y}_{n_i} = \alpha + \beta x_i + \bar{\epsilon}_{n_i} \tag{74}$$

where

$$\begin{aligned} E(\bar{\epsilon}_{n_i}^2) &= \frac{\mathbf{S}_i^2}{n_i} \cdot \frac{N_i - n_i}{N_i} \\ &= \frac{\gamma x_i}{n_i} \cdot \frac{N_i - n_i}{N_i - 1} \\ &= \frac{\gamma}{w_i'} \end{aligned} \tag{75}$$

Also, from (74),

$$\sum_{i=1}^{k} w_i' \bar{y}_{n_i} = \alpha \sum_{i=1}^{k} w_i' + \beta \sum_{i=1}^{k} w_i' x_i + \sum_{i=1}^{k} w_i' \bar{\epsilon}_{n_i}$$

whence, dividing by W', we have

$$\bar{y}_w = \alpha + \beta \bar{x}_w + \bar{\epsilon}_w \tag{76}$$

where

$$\bar{\epsilon}_w = \frac{\sum_{i=1}^{k} w_i' \bar{\epsilon}_{n_i}}{W'} \tag{77}$$

Further, from (65),

$$\hat{\beta} = \frac{1}{s_{wx}^{2}} \left[\frac{\sum\limits_{i=1}^{k} w_i' x_i \bar{y}_{n_i}}{W'} - \bar{x}_w \bar{y}_w \right]$$

$$= \frac{1}{s_{wx}^{2}} \left[\frac{\sum\limits_{i=1}^{k} w_i' x_i \bar{y}_{n_i}}{W'} - \bar{x}_w \cdot \frac{\sum\limits_{i=1}^{k} w_i' \bar{y}_{n_i}}{W'} \right]$$

$$= \frac{1}{W' s_{wx}^{2}} \left[\sum_{i=1}^{k} w_i' (x_i - \bar{x}_w)\, \bar{y}_{n_i} \right]$$

$$= \frac{1}{W' s_{wx}^{2}} \left[\sum_{i=1}^{k} w_i' (x_i - \bar{x}_w)(\alpha + \beta x_i + \bar{\epsilon}_{n_i}) \right]$$

$$= \frac{1}{W' s_{wx}^{2}} \left[\sum_{i=1}^{k} \beta w_i' (x_i - \bar{x}_w)^2 + \sum_{i=1}^{k} w_i' (x_i - \bar{x}_w)\, \bar{\epsilon}_{n_i} \right]$$

$$= \beta + \frac{1}{W' s_{wx}^{2}} \sum_{i=1}^{k} w_i' (x_i - \bar{x}_w)\, \bar{\epsilon}_{n_i} \tag{78}$$

On substituting for y_{ij} from (1), for $\bar{y}_w$ from (76) and for $\hat{\beta}$ from (78) in (72), we obtain

$$\mathbf{Q} = \sum_{i=1}^{k} \sum_{j}^{n_i} \frac{w_i'}{n_i} \Bigg\{ \alpha + \beta x_i + \epsilon_{ij} - \alpha - \beta \bar{x}_w - \bar{\epsilon}_w$$

$$- (x_i - \bar{x}_w) \left(\beta + \frac{1}{W' s_{wx}^{2}} \sum_{i=1}^{k} w_i' (x_i - \bar{x}_w)\, \bar{\epsilon}_{n_i} \right) \Bigg\}^2$$

$$= \sum_{i=1}^{k} \sum_{j}^{n_i} \frac{w_i'}{n_i} \Bigg\{ (\epsilon_{ij} - \bar{\epsilon}_w)$$

$$- (x_i - \bar{x}_w) \cdot \frac{1}{W' s_{wx}^{2}} \sum_{i=1}^{k} w_i' (x_i - \bar{x}_w)\, \bar{\epsilon}_{n_i} \Bigg\}^2$$

$$= \sum_{i=1}^{k} \sum_{j}^{n_i} \frac{w_i'}{n_i} (\epsilon_{ij} - \bar{\epsilon}_w)^2$$

$$+ \sum_{i=1}^{k} \sum_{j}^{n_i} \frac{w_i'}{n_i} \cdot \frac{1}{W'^2 s_{wx}^{\;4}} (x_i - \bar{x}_w)^2$$

$$\times \left\{ \sum_{t=1}^{k} w_t' (x_t - \bar{x}_w) \bar{\epsilon}_{nt} \right\}^2$$

$$-2 \sum_{i=1}^{k} \sum_{j}^{n_i} \frac{w_i'}{n_i} \cdot \frac{1}{W' s_{wx}^{\;2}} (x_i - \bar{x}_w) (\epsilon_{ij} - \bar{\epsilon}_w)$$

$$\times \left\{ \sum_{t=1}^{k} w_t' (x_t - \bar{x}_w) \bar{\epsilon}_{nt} \right\}$$

$$= \sum_{i=1}^{k} \sum_{j}^{n_i} \frac{w_i'}{n_i} \epsilon_{ij}^{\;2} + \bar{\epsilon}_w^{\;2} \sum_{i=1}^{k} \sum_{j}^{n_i} \frac{w_i'}{n_i} - 2\, \bar{\epsilon}_w \sum_{i=1}^{k} \sum_{j}^{n_i} \frac{w_i' \epsilon_{ij}}{n_i}$$

$$+ \frac{1}{W' s_{wx}^{\;2}} \left\{ \sum_{t=1}^{k} w_t' (x_t - \bar{x}_w) \bar{\epsilon}_{nt} \right\}^2$$

$$- \frac{2}{W' s_{wx}^{\;2}} \left\{ \sum_{t=1}^{k} w_t' (x_t - \bar{x}_w) \bar{\epsilon}_{nt} \right\}^2$$

$$= \sum_{i=1}^{k} \sum_{j}^{n_i} \frac{w_i'}{n_i} \epsilon_{ij}^{\;2} - W' \bar{\epsilon}_w^{\;2}$$

$$- \frac{1}{W' s_{wx}^{\;2}} \left\{ \sum_{t=1}^{k} w_t' (x_t - \bar{x}_w) \bar{\epsilon}_{nt} \right\}^2$$

$$= \sum_{i=1}^{k} \sum_{j}^{n_i} \frac{w_i'}{n_i} \epsilon_{ij}^{\;2} - \frac{1}{W'} \left\{ \sum_{i=1}^{k} w_i'^2 \bar{\epsilon}_{ni}^{\;2} + \sum_{i \neq h=1}^{k} w_i' w_h' \bar{\epsilon}_{ni} \bar{\epsilon}_{nh} \right\}$$

$$- \frac{1}{W' s_{wx}^{\;2}} \left\{ \sum_{i=1}^{k} w_i'^2 \bar{\epsilon}_{ni}^{\;2} (x_i - \bar{x}_w)^2 \right.$$

$$\left. + \sum_{i \neq h=1}^{k} w_i' w_h' \bar{\epsilon}_{ni} \bar{\epsilon}_{nh} (x_i - \bar{x}_w)(x_h - \bar{x}_w) \right\} \quad (79)$$

On taking expectations and noting that

$$E(\bar{\epsilon}_{n_i}^{\,2}) = \frac{\gamma}{w_i'}$$

and

$$E(\bar{\epsilon}_{n_i}\bar{\epsilon}_{n_h}) = 0$$

we have

$$E(\mathbf{Q}) = \gamma\left[\sum_{i=1}^{k} \frac{n_i(N_i - 1)}{N_i - n_i} - 2\right] \tag{80}$$

It follows that an unbiased estimate of γ is provided by

$$\text{Est. } \gamma = \frac{\mathbf{Q}}{\left(\sum_{i=1}^{k} \frac{n_i(N_i - 1)}{N_i - n_i} - 2\right)} \tag{81}$$

Now, let

$$s_{wy}^{2} = \frac{\sum_{i=1}^{k} \frac{w_i'}{n_i} \sum_{j}^{n_i} (y_{ij} - \bar{y}_w)^2}{\sum_{i=1}^{k} w_i'} \tag{82}$$

and

$$r = \frac{\sum_{i=1}^{k} w_i'(x_i - \bar{x}_w)\,\bar{y}_{n_i}}{s_{wx}s_{wy}W'} \tag{83}$$

then **Q** can be expressed as

$$\mathbf{Q} = W's_{wy}^{2}(1 - r^2) \tag{84}$$

and we have

$$\text{Est. } V(Y_{wl}) = \frac{N^2 s_{wy}^{2}(1 - r^2)\left[1 + \frac{(\bar{x}_N - \bar{x}_w)^2}{s_{wx}^{2}}\right]}{\sum_{i=1}^{k} \frac{n_i(N_i - 1)}{N_i - n_i} - 2} \tag{85}$$

When sampling is carried out with replacement and

$$V(y \mid x) = \gamma$$

then

$$w_i = \frac{n_i}{\gamma}, \; \bar{y}_w = \bar{y}_n$$

$$s_{wy}^{\,2} = \frac{\sum_{i=1}^{k} \sum_{j}^{n_i} (y_{ij} - \bar{y}_n)^2}{n}$$

$$r = \frac{\sum_{i=1}^{k} \sum_{j}^{n_i} y_{ij}\,(x_i - \bar{x}_n)}{n s_{wx} s_{wy}}$$

$$\text{Est. } \gamma = \frac{\mathbf{Q}}{n-2} = \frac{n s_{wy}^{\,2}\,(1-r^2)}{n-2}$$

and we have, as previously,

$$\text{Est. } V(Y_l) = \frac{N^2 s_{wy}^{\,2}\,(1-r^2)}{n-2}\left[1 + \frac{(\bar{x}_N - \bar{x}_n)^2}{s_{wx}^{\,2}}\right] \tag{86}$$

5.7 Comparison of Weighted with Simple Regression

The sampling variance of the weighted regression estimate, to terms in $1/n$, is obtained from (66), being given by

$$V(Y_{wl}) = \frac{N^2}{W} \tag{87}$$

where

$$W = \sum_{i=1}^{k} w_i$$

$$= \sum_{i=1}^{k} \frac{n_i}{S_i^2} \cdot \frac{N_i}{N_i - n_i}$$

while that of the simple regression estimate to the same degree of accuracy is from (22) given by

$$V(Y_l) = \frac{N^2 \gamma}{n} \tag{88}$$

However, (87) and (88) are not directly comparable. To make them comparable, we shall suppose that sampling is carried out with replacement, so that we may take

$$\frac{N_i}{N_i - n_i} = 1 \,.$$

Equation (87) then becomes

$$V\,(Y_{wl}) = \frac{N^2}{\sum\limits_{i=1}^{k} \dfrac{n_i}{\sigma_i^2}} \tag{89}$$

where

$$\sigma_i^2 = \frac{\sum\limits_{j=1}^{N_i} (y_{ij} - \bar{y}_{N_i})^2}{N_i}$$

and

$$\frac{V\,(Y_{wl})}{V\,(Y_1)} = \frac{n}{\gamma \sum\limits_{i=1}^{k} \dfrac{n_i}{\sigma_i^2}} \tag{90}$$

Now, let

$$\sigma_i^2 = \gamma + \delta\,(\sigma_i^2) \tag{91}$$

so that

$$E\,\{\delta\,(\sigma_i^2)\} = 0, \quad E\,\{\delta\,(\sigma_i^2)\}^2 = V\,(\sigma_i^2)$$

and

$$\sum_{i=1}^{k} \frac{n_i}{\sigma_i^2} = \sum_{i=1}^{k} \frac{n_i}{\gamma\left(1 + \dfrac{\delta\,(\sigma_i^2)}{\gamma}\right)}$$

$$= \sum_{i=1}^{k} \frac{n_i}{\gamma}\left(1 - \frac{\delta\,(\sigma_i^2)}{\gamma} + \frac{\{\delta\,(\sigma_i^2)\}^2}{\gamma^2} - \cdots\right) \tag{92}$$

provided

$$\left|\frac{\delta\,(\sigma_i^2)}{\gamma}\right| < 1$$

Taking expectations, we obtain

$$E\left(\sum_{i=1}^{k}\frac{n_i}{\sigma_i^2}\right)\cong\frac{n}{\gamma}+\frac{n}{\gamma}\,C^2_{\sigma_i^2} \qquad (93)$$

where $C^2_{\sigma_i^2}$ is the square of the coefficient of variation of σ_i^2. The average value of (90) is, therefore, approximated by

$$E\left\{\frac{V(Y_{wl})}{V(Y_l)}\right\}\cong\frac{1}{1+C^2_{\sigma_i^2}} \qquad (94)$$

The result is due to Cochran (1942). It shows what is indeed obvious, that if the σ_i^2's do not change very greatly, simple regression may be used without appreciable loss of precision. If, on the other hand, σ_i^2's vary very considerably, the use of the simple regression will lead to loss of efficiency.

Example 5.1

Table 5.1 summarizes the data for a simple random sample of 64 villages drawn from the total of 319 villages referred to in Example 4.2. Assuming that villages within a class are of the same size, equal to the mean value per village in that class,

TABLE 5.1

Summary of Data Relating to Agricultural Area (x) and Number of Livestock (y) in a Simple Random Sample of 64 Villages Selected from the Population in Table 4.1

Serial No. of Class	Area of Village (x_i)	n_i	$\sum_j^{n_i} y_{ij}$	$\sum_j^{n_i} y_{ij}^2$
1	63·73	2	16	256
2	155·33	5	333	26895
3	245·68	18	1810	281314
4	344·40	16	1991	330113
5	491·56	13	1815	287079
6	767·49	10	2352	605510
Total		64	8317	1531167

use the method of linear regression to estimate the livestock population and its variance under each of the following two assumptions:

I. $V(y \mid x) = \gamma x$

and, II. $V(y \mid x) =$ constant, say γ; and sampling within each class is with replacement.

Method I

The relevant formulæ for the regression estimate of the population total and its variance are given by (63) and (85). To evaluate these we require the values of $\bar{x}_w$, $\bar{y}_w$, $s_{wx}{}^2$, $s_{wy}{}^2$ and the estimate of β. The calculations leading to these values are given in Table 5.2. From this table we obtain

$$\bar{x}_w = \frac{\text{Total of col. 5}}{\text{Total of col. 7}}$$

$$= 294 \cdot 24$$

$$\bar{y}_w = \frac{\text{Total of col. 9}}{\text{Total of col. 7}}$$

$$= 102 \cdot 72$$

$$s_{wx}{}^2 = \frac{\sum_{i=1}^{k} w_i{}' (x_i - \bar{x}_w)^2}{W'}$$

$$= \frac{\text{Total of col. 16}}{\text{Total of col. 7}}$$

$$= \frac{7952}{\cdot 27297}$$

$$= 29131$$

$$s_{wx} = 170 \cdot 7.$$

$$s_{wy}{}^2 = \frac{\sum_{i=1}^{k} \frac{w_i{}'}{n_i} \sum_{j}^{n_i} y_{ij}{}^2 - W'\bar{y}_w{}^2}{W'}$$

$$= \frac{4646\cdot4 - 2880\cdot2}{\cdot27297}$$

$$= \frac{1766\cdot2}{\cdot27297}$$

$$= 6470\cdot3$$

$$s_{wy} = 80\cdot44$$

$$r = \frac{\sum_{i=1}^{k} w_i' \bar{y}_{n_i} (x_i - \bar{x}_w)}{s_{wx} s_{wy} W'}$$

$$= \frac{2314\cdot6}{(13731)\ (\cdot27297)}$$

$$= \frac{8479\cdot3}{13731}$$

$$= \cdot6175$$

$$\hat{\beta} = \frac{\sum_{i=1}^{k} w_i' \bar{y}_{n_i} (x_i - \bar{x}_w)}{W' s_{wx}{}^2}$$

$$= \frac{2314\cdot55}{7952}$$

$$= \cdot2911$$

Hence on substituting in (63), we have

$$\begin{aligned} Y_{wt} &= 319\ [\bar{y}_w + \hat{\beta}\ (\bar{x}_N - \bar{x}_w)] \\ &= 319\ [102\cdot7 + \cdot2911\ (367\cdot5 - 294\cdot2)] \\ &= 319\ [102\cdot7 + \cdot2911\ (73\cdot3)] \\ &= 319\ [124\cdot0] \\ &= 39556 \end{aligned}$$

Also, on substituting in (85), we get

$$\text{Est. } V\,(Y_{wt}) = \frac{(319)^2\,(6470\cdot3)\,(1 - \cdot3813)}{78\cdot3191}\left[1 + \frac{(73\cdot3)^2}{29131}\right]$$

$$= 101761 \times 51{\cdot}1136\ [1 + {\cdot}1844]$$

$$= 101761 \times 60{\cdot}5389$$

$$\%\ S.E.\ Y_{wl} = \frac{100\ \sqrt{60{\cdot}54}}{124} = \frac{778}{124}$$

$$= 6{\cdot}27$$

Method II

The relevant formulæ are given by (19) and (86) respectively. We have

$$\bar{y}_n = \frac{8317}{64}$$

$$= 129{\cdot}95$$

$$\bar{x}_n = \frac{24902}{64}$$

$$= 389{\cdot}09$$

$$\hat{\beta} = \frac{\sum\limits_{i=1}^{k} \sum\limits_{j}^{n_i} y_{ij}\,(x_i - \bar{x}_n)}{\sum\limits_{i=1}^{k} n_i\,(x_i - \bar{x}_n)^2}$$

$$= \frac{644382}{2455455}$$

$$= 0{\cdot}2624$$

Hence on substituting in (19), we have

$$Y_l = 319\ [129{\cdot}95 + 0{\cdot}2624\ (367{\cdot}5 - 389{\cdot}1)]$$

$$= 319\ (124{\cdot}28)$$

$$= 39645$$

Again

$$s_{wy}^{\ 2} = \frac{\sum\limits_{i=1}^{k} \sum\limits_{j}^{n_i} y_{ij}^{\ 2} - n\bar{y}_n^{\ 2}}{n}$$

$$= \frac{1531167 - 1080768}{64}$$

$$= 7037{\cdot}5$$

$$s_{wx}^2 = \frac{\sum_{i=1}^{k} n_i x_i^2 - n\bar{x}_n^2}{n}$$

$$= \frac{2455454 \cdot 7}{64}$$

$$= 38366 \cdot 48$$

and

$$s_{wyx} = \frac{\sum_{i=1}^{k} n_i \bar{y}_{n_i} (x_i - \bar{x}_n)}{n}$$

$$= \frac{644382 \cdot 12}{64}$$

$$= 10068 \cdot 47$$

$$(1 - r^2)\, s_{wy}^2 = s_{wy}^2 - r^2 s_{wy}^2 = s_{wy}^2 - \frac{s_{wyx}^2}{s_{wx}^2}$$

$$= 7037 \cdot 5 - \frac{(10068 \cdot 47)^2}{38366 \cdot 48}$$

$$= 4395 \cdot 2$$

Hence, on substituting in (86), we get

$$V(Y_l) = 319^2 \cdot \frac{4395 \cdot 2}{62} \cdot \left(1 + \frac{(367 \cdot 53 - 389 \cdot 09)^2}{38366 \cdot 5}\right)$$

$$= 319^2 \times 70 \cdot 89 \times 1 \cdot 01212$$

$$= 101761 \times 71 \cdot 75$$

$$\%\ S.E.\ Y_l = \frac{100\sqrt{71 \cdot 75}}{124 \cdot 3} = 6 \cdot 81$$

Compared to the value of 60·54 of the sampling variance of the estimated mean obtained by Method I, Method II is seen to give a value of 71·75, which is larger by about 20%. This must be traced to the rather large variability among σ_i^2's, *vide* Table 4.1.

TABLE 5.2

Computations Leading to the Values of the Regression Estimate of the Livestock Population and its Variance

Serial No. of Class	n_i	N_i	$n_i(N_i-1)$	N_i-n_i	$w_i'x_i$	x_i	w_i'	$\bar{y}_{n_i}$
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1	2	11	20	9	2·2222	63·73	·03487	8·00
2	5	48	235	43	5·4651	155·33	·03518	66·60
3	18	84	1494	66	22·6364	245·68	·09214	100·56
4	16	60	944	44	21·4545	344·40	·06230	124·44
5	13	77	988	64	15·4375	491·56	·03141	139·62
6	10	39	380	29	13·1034	767·49	·01707	235·20
Total	64	319			80·3191		·27297	

Serial No. of Class	$w_i'\bar{y}_{n_i}$	$\sum_{j}^{n_i} y_{ij}^2$	$\dfrac{(10)}{n_i}$	$(11)\cdot w_i'$	$x_i-\bar{x}_w$	$(9)\cdot(13)$	$(x_i-\bar{x}_w)^2$	$w_i'(x_i-\bar{x}_w)^2$
	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)
1	0·2790	256	128	4·5	−230·51	− 64·31	53135	1853
2	2·3430	26895	5379	189·2	−138·91	−325·47	19296	679
3	9·2656	281314	15629	1440·1	− 48·56	−449·94	2358	217
4	7·7526	330113	20632	1285·4	50·16	388·87	2516	157
5	4·3855	287079	22083	693·6	197·32	865·35	38935	1223
6	4·0149	605510	60551	1033·6	473·25	1900·05	223966	3823
Total	28·0406			4646·4		2314·55		7952

5.8 Comparison of Simple Regression with the Ratio and the Mean per Sampling Unit Estimates

The sampling variance of the simple regression estimate of the population total to terms in $1/n$ has been shown to be

$$V(Y_l) = N^2\sigma_y^2\,\frac{(1-\rho^2)}{n} \tag{95}$$

The sampling variance of the ratio estimate under comparable conditions of sampling with replacement is obtained by dropping the finite multiplier from (28) in the previous Chapter, and will, therefore, be

$$V(Y_R) = \frac{N^2}{n}(\sigma_y^2 - 2R_N\rho\sigma_y\sigma_x + R_N^2\sigma_x^2) \tag{96}$$

while that of the mean per sampling unit estimate is given by

$$V(N\bar{y}_n) = \frac{N^2\sigma_y^2}{n} \tag{97}$$

Comparing first the simple regression with the mean per sampling unit estimate, we notice that the regression estimate is always more accurate than the arithmetic mean estimate. Comparing next the simple regression with the ratio estimate, we observe that the former is more accurate than the latter if

$$\sigma_y^2 - 2R_N\rho\sigma_y\sigma_x + R_N^2\sigma_x^2 > \sigma_y^2 - \sigma_y^2\rho^2$$

i.e., if

$$\rho^2\sigma_y^2 - 2R_N\rho\sigma_y\sigma_x + R_N^2\sigma_x^2 > 0$$

i.e., if

$$(\rho\sigma_y - R_N\sigma_x)^2 > 0 \tag{98}$$

which is always true, unless

$$R_N = \rho\,\frac{\sigma_y}{\sigma_x}$$

Hence the regression estimate is always more accurate than the ratio estimate unless the regression of y on x is a straight line passing through the origin, in which case the two estimates will have equal variance.

The data for 319 villages referred to in Example 4.2 provide the material for the comparison of the two methods. Substituting for values of σ_y^2, ρ, σ_x^2 and R_N from col. 8 of Table 4.1 in formulæ (95) and (96), we obtain

$$\frac{V(Y_l)}{V(Y_R)} = \frac{\sigma_y^2(1-\rho^2)}{\sigma_y^2 - 2R_N\rho\sigma_y\sigma_x + R_N^2\sigma_x^2}$$

$$= \frac{4416 \cdot 3}{4416 \cdot 6}$$

$$\cong 1$$

There is thus little to choose between the simple regression and the ratio methods of estimation, since the regression of the number of livestock on agricultural area is almost a straight line passing through the origin.

5.9 Comparison of Simple Regression with Stratified Sampling

The regression method of estimation achieves the same purpose as stratification by size of the sampling unit, namely, to eliminate the effect of variation in the size of the sampling unit from the standard error of the estimated character. A comparison of the two methods is, therefore, of interest.

We have seen that the sampling variance of the estimated total in stratified sampling is given by

$$N^2 \sum_{i=1}^{k} p_i^2 \frac{N_i - n_i}{N_i} \cdot \frac{S_i^2}{n_i}$$

This, however, is not the appropriate variance to compare with the variance of the simple regression estimate. The appropriate variance for comparison would be the variance of a simple random sample whose units can be classified by strata, and which can then be treated as if it were selected by the method of stratified sampling, the sampling within each stratum to be with replacement. This is directly obtained from (87) of Chapter III, being given by

$$\frac{N^2}{n}\left\{\left(1-\frac{1}{n}\right)\sum_{i=1}^{k} p_i\sigma_i^2 + \frac{1}{n}\sum_{i=1}^{k}\sigma_i^2\right\} \tag{99}$$

In simple regression we further assume that the true regression is linear, with constant residual variance. If the σ_i^2's, therefore, are approximately constant, say equal to σ_w^2. the expression (99) will reduce to

$$\frac{N^2\sigma_w^2}{n}\left(1 + \frac{k-1}{n}\right) \tag{100}$$

This is the variance which is comparable with the average variance of the simple regression estimate, namely,

$$\frac{N^2\sigma_y^2}{n}(1-\rho^2)\left(1 + \frac{1}{n}\right) \tag{101}$$

Now the value of $\sigma_y^2 (1 - \rho^2)$, as it represents the residual variance about the regression straight line, can never be less than σ_w^2. It follows therefore that the stratified sample will, in general, furnish a more accurate estimate than the simple regression method. The relationship between y and x is also not always found to be linear in practice, in which case the efficiency of the regression estimate is further reduced. For, while stratified sampling with suitably chosen strata can take care of any type of relationship, the regression estimate can eliminate only the effects of the linear component of the relationship. Stratified sampling has an added advantage, in that the estimate for this procedure is an unbiased estimate for any type of relationship between x and y, and for any size of sample. It would, therefore, seem that provided the population is divided into an adequate number of strata, so as to make σ_w^2 small, we may expect stratified sampling to be superior to the regression method of estimation under most practical conditions.

5.10 Double Sampling

The regression estimate of the population mean of y presupposes that the population mean of x, namely $\bar{x}_N$, is known. However, $\bar{x}_N$ is not always known although in many inquiries it is possible to estimate it from a second sample of the population without appreciably adding to the cost of the inquiry. The procedure is known as *double sampling*. In this section, we shall give the form of the regression estimate in double sampling and its sampling variance.

Let, as previously, the population be divided into k classes with N_i units having the value x_i each ($i = 1, 2, \ldots, k$). We shall suppose that n denotes a simple random sample of N on which both y and x are observed, and that in repeated samples of $n, n_1, n_2, \ldots, n_k$ units are drawn from their respective classes with replacement. Further, we shall suppose that a second random sample Q' is drawn from the remaining $N - n$, say N', units of the population and that only x is observed on this sample.

Now if $\bar{x}_N$ were known, then the simple regression estimate of $\bar{y}_N$ would be given by (19), namely,

$$\bar{y}_l = \bar{y}_n + b\,(\bar{x}_N - \bar{x}_n) \tag{102}$$

where

$$b = \frac{\sum\limits_{i=1}^{k} n_i \bar{y}_{n_i} (x_i - \bar{x}_n)}{\sum\limits_{i=1}^{k} n_i (x_i - \bar{x}_n)^2}$$

This can be rewritten as

$$\bar{y}_l = \bar{y}_n + b\,\frac{N'}{N}\,(\bar{x}_{N'} - \bar{x}_n) \tag{103}$$

Since Q' is a random sample of N', $\bar{x}_{Q'}$ provides the best unbiased linear estimate of $\bar{x}_{N'}$. Hence

$$\bar{y}_{ds} = \bar{y}_n + b\,\frac{N'}{N}\,(\bar{x}_{Q'} - \bar{x}_n) \tag{104}$$

where $\bar{y}_{ds}$ denotes the estimate of $\bar{y}_N$ based on a double sample of n and Q'.

It is easily shown that $\bar{y}_{ds}$ is an unbiased estimate of $\bar{y}_N$. We write

$$E\,(\bar{y}_{ds} \mid n_1, n_2, \ldots, n_k, Q') = E\,(\bar{y}_n) + \frac{N'}{N}\; E\,\{b\,(\bar{x}_{Q'} - \bar{x}_n)\} \tag{105}$$

Now from (31) and (32), we obtain

$$E\,(\bar{y}_n) = \alpha + \beta \bar{x}_n$$

and

$$\tag{106}$$

$$E\,(b) = \beta$$

$$\tag{107}$$

Also

$$E\,(b\bar{x}_{Q'}) = E\left[\frac{\left\{\sum_{i=1}^{k} n_i\bar{y}_{n_i}\,(x_i - \bar{x}_n)\right\}\left\{\sum_{i=1}^{k} Q_i'x_i\right\}}{\left\{\sum_{i=1}^{k} n_i\,(x_i - \bar{x}_n)^2\right\}Q'}\right]$$

where Q_i' represents the number of units in the i-th class in the sample Q',

$$= \frac{E\left\{\sum_{i=1}^{k} n_i\,\frac{Q_i'}{Q'}\,\bar{y}_{n_i}x_i\,(x_i - \bar{x}_n)\right\} + E\left\{\sum_{i\neq j=1}^{k} n_i\bar{y}_{n_i}\,(x_i - \bar{x}_n)\,\frac{Q_j'}{Q'}\,x_j\right\}}{\sum_{i=1}^{k} n_i\,(x_i - \bar{x}_n)^2}$$

$$= \frac{\sum_{i=1}^{k} n_i\,\frac{N_i'}{N'}\,\bar{y}_{N_i}x_i\,(x_i - \bar{x}_n) + \sum_{i\neq j=1}^{k} n_i\bar{y}_{N_i}\,(x_i - \bar{x}_n)\,\frac{N_j'}{N'}\,x_j}{\sum_{i=1}^{k} n_i\,(x_i - \bar{x}_n)^2}$$

Substituting $\bar{y}_{N_i} = \alpha + \beta x_i$, we get

$$E\,(b\bar{x}_{Q'}) = \left\{\frac{\sum_{i=1}^{k} n_i\,(\alpha + \beta x_i)\,(x_i - \bar{x}_n)}{\sum_{i=1}^{k} n_i\,(x_i - \bar{x}_n)^2}\right\}\cdot\left\{\frac{\sum_{i=1}^{k} N_i'x_i}{N'}\right\}$$

$$= \beta\bar{x}_{N'} \tag{108}$$

On substituting from (106), (107) and (108) in (105), we get

$$E\,(\bar{y}_{ds}\mid n_1, n_2, \ldots, n_k, Q') = \alpha + \beta\bar{x}_n + \frac{N'}{N}\,\beta\,(\bar{x}_{N'} - \bar{x}_n)$$

$$= \alpha + \beta\bar{x}_n + \beta\,(\bar{x}_N - \bar{x}_n)$$

$$= \alpha + \beta\bar{x}_N$$

$$= \bar{y}_N \tag{109}$$

thus showing that $\bar{y}_{ds}$ is an unbiased estimate of $\bar{y}_N$.

To obtain the variance, we write

$$V(\bar{y}_{ds}) = E\left[\left\{\bar{y}_n + b\,\frac{N'}{N}\,(\bar{x}_{Q'} - \bar{x}_n) - \bar{y}_N\right\}^2\right]$$

$$= E(\bar{y}_n^2) + \frac{N'^2}{N^2}\,E\left\{b^2\,(\bar{x}_{Q'} - \bar{x}_n)^2\right\}$$

$$+ 2\,\frac{N'}{N}\,E\left\{\bar{y}_n b\,(\bar{x}_{Q'} - \bar{x}_n)\right\} - \bar{y}_N^2 \qquad (110)$$

Now

$$E(\bar{y}_n^2) = E\left(\frac{1}{n}\sum_{i=1}^{k} n_i\bar{y}_{n_i}\right)^2$$

$$= \frac{1}{n^2}\,E\left\{\sum_{i=1}^{k} n_i^2\bar{y}_{n_i}^2 + \sum_{i\neq j=1}^{k} n_i n_j \bar{y}_{n_i}\bar{y}_{n_j}\right\}$$

$$= \frac{1}{n^2}\left\{\sum_{i=1}^{k} n_i^2\left(\frac{\gamma}{n_i} + \bar{y}_{N_i}^2\right) + \sum_{i\neq j=1}^{k} n_i n_j \bar{y}_{N_i}\bar{y}_{N_j}\right\}$$

$$= \frac{\gamma}{n} + \left(\frac{1}{n}\sum_{i=1}^{k} n_i\bar{y}_{N_i}\right)^2$$

$$= \frac{\gamma}{n} + (\alpha + \beta\bar{x}_n)^2 \qquad (111)$$

Next

$$E(b^2) = \beta^2 + E(b - \beta)^2$$

On substituting from (32) and taking expectations, we get

$$E(b^2) = \beta^2 + \frac{\gamma}{\sum\limits_{i=1}^{k} n_i\,(x_i - \bar{x}_n)^2} \qquad (112)$$

and

$$E(\bar{y}_n b) = E\left[\left(\frac{1}{n}\sum_{i=1}^{k} n_i\bar{y}_{n_i}\right)\left\{\frac{\sum\limits_{i=1}^{k} n_i\bar{y}_{n_i}\,(x_i - \bar{x}_n)}{\sum\limits_{i=1}^{k} n_i\,(x_i - \bar{x}_n)^2}\right\}\right]$$

$$= \frac{1}{n \sum_{i=1}^{k} n_i (x_i - \bar{x}_n)^2} E \left\{ \sum_{i=1}^{k} n_i^2 \bar{y}_{n_i}^2 (x_i - \bar{x}_n) + \sum_{i \neq j=1}^{k} n_i n_j \bar{y}_{n_i} \bar{y}_{n_j} (x_i - \bar{x}_n) \right\}$$

$$= \frac{1}{n \sum_{i=1}^{k} n_i (x_i - \bar{x}_n)^2} \cdot \left\{ \sum_{i=1}^{k} n_i^2 \left(\frac{\gamma}{n_i} + \bar{y}_{N_i}^2 \right) (x_i - \bar{x}_n) + \sum_{i \neq j=1}^{k} n_i n_j \bar{y}_{N_i} \bar{y}_{N_j} (x_i - \bar{x}_n) \right\}$$

$$= \frac{1}{n \sum_{i=1}^{k} n_i (x_i - \bar{x}_n)^2} \left\{ \left(\sum_{i=1}^{k} n_i \bar{y}_{N_i} \right) \times \left(\sum_{i=1}^{k} n_i \bar{y}_{N_i} (x_i - \bar{x}_n) \right) \right\}$$

$$= \frac{1}{n \sum_{i=1}^{k} n_i (x_i - \bar{x}_n)^2} \left[\left\{ \sum_{i=1}^{k} n_i (\alpha + \beta x_i) \right\} \times \left\{ \sum_{i=1}^{k} n_i (\alpha + \beta x_i) (x_i - \bar{x}_n) \right\} \right]$$

$$= (\alpha + \beta \bar{x}_n) \cdot \beta$$

$$= E(\bar{y}_n) \cdot E(b) \tag{113}$$

Substituting from (111), (112) and (113) in (110), we therefore have

$$V(\bar{y}_{ds}) = \frac{\gamma}{n} + (\alpha + \beta \bar{x}_n)^2 + \frac{N'^2}{N^2} \left\{ \beta^2 + \frac{\gamma}{\sum_{i=1}^{k} n_i (x_i - \bar{x}_n)^2} \right\} \times E(\bar{x}_{Q'} - \bar{x}_n)^2$$

$$+ 2 \frac{N'}{N} (\alpha + \beta \bar{x}_n) \beta \cdot E(\bar{x}_{Q'} - \bar{x}_n) - \bar{y}_N^2 \tag{114}$$

Lastly,

$$
\begin{aligned}
E(\bar{x}_{Q'} - \bar{x}_n)^2 &= E(\bar{x}_{Q'} - \bar{x}_{N'} + \bar{x}_{N'} - \bar{x}_n)^2 \\
&= E\{(\bar{x}_{Q'} - \bar{x}_{N'})^2 + (\bar{x}_{N'} - \bar{x}_n)^2\} \\
&= E(\bar{x}_{Q'} - \bar{x}_{N'})^2 + \frac{N^2}{N'^2}(\bar{x}_N - \bar{x}_n)^2 \\
&= \left(\frac{1}{Q'} - \frac{1}{N'}\right) \mathrm{S}^2_{x|N'} + \frac{N^2}{N'^2}(\bar{x}_N - \bar{x}_n)^2 \qquad (115)
\end{aligned}
$$

where

$$
\mathrm{S}^2_{x|N'} = \frac{1}{N'-1} \sum_{i=1}^{N'} (x_i - \bar{x}_{N'})^2
$$

Also

$$
\begin{aligned}
\frac{N'}{N} E(\bar{x}_{Q'} - \bar{x}_n) &= \frac{N'}{N}(\bar{x}_{N'} - \bar{x}_n) \\
&= \bar{x}_N - \bar{x}_n \qquad (116)
\end{aligned}
$$

On substituting from (25), (115) and (116) in (114) and simplifying, we get

$$
V(\bar{y}_{ds}) = \gamma\left[\frac{1}{n} + \frac{(\bar{x}_N - \bar{x}_n)^2}{\sum\limits_{i=1}^{k} n_i (x_i - \bar{x}_n)^2}\right] + \frac{N'^2}{N^2}\left(\frac{1}{Q'} - \frac{1}{N'}\right)\mathrm{S}^2_{x|N'} \times \left\{\frac{\gamma}{\sum\limits_{i=1}^{k} n_i (x_i - \bar{x}_n)^2} + \beta^2\right\} \qquad (117)
$$

For N and N' large, (117) reduces to

$$
V(\bar{y}_{ds}) \cong \gamma\left[\frac{1}{n} + \frac{(\bar{x}_N - \bar{x}_n)^2}{\sum\limits_{i=1}^{k} n_i (x_i - \bar{x}_n)^2}\right] + \frac{\mathrm{S}_x^{\,2}}{Q'}\left\{\frac{\gamma}{\sum\limits_{i=1}^{k} n_i (x_i - \bar{x}_n)^2} + \beta^2\right\} \qquad (118)
$$

It will be seen that the variance is composed of two terms. The first term is clearly the variance of the simple regression estimate when $\bar{x}_N$ is known, while the second represents the increase in the variance of the estimate when $\bar{x}_N$ is determined from a sample. When $Q' = N'$ or, in other words, when $\bar{x}_N$ is known, we are left with the variance of the simple regression estimate.

The variance of the estimate $\bar{y}_{ds}$ is seen to depend upon the x's in the sample. Its average value is easily deduced following the method adopted in Section 5.4. Thus, for large N and ignoring terms of order higher than $1/n^2$, we have from (42),

$$E\left\{\frac{(\bar{x}_N - \bar{x}_n)^2}{\sum\limits_{i=1}^{k} n_i (x_i - \bar{x}_n)^2}\right\} \cong \frac{1}{n^2} \tag{119}$$

Also

$$E\left\{\frac{S_x^2}{Q'} \cdot \frac{\gamma}{\sum\limits_{i=1}^{k} n_i (x_i - \bar{x}_n)^2}\right\} \cong \frac{\gamma}{Q'n} \tag{120}$$

On substituting from (119) and (120) in (118), we get

$$E\,\{V\,(\bar{y}_{ds})\} \cong \frac{\gamma}{n}\left[1 + \frac{1}{n} + \frac{1}{Q'}\right] + \frac{\rho^2 S_y^2}{Q'} \tag{121}$$

For large N, an alternative and more efficient estimate than the one given by (104) can be formed. This is defined by

$$\bar{y}'_{ds} = \bar{y}_n + b\,(\bar{x}_{Q'+n} - \bar{x}_n) \tag{122}$$

It is easily shown that this is an unbiased estimate of the population mean, for,

$$\begin{aligned} E\,(\bar{y}'_{ds}) &= (\alpha + \beta\bar{x}_n) + \beta E\,(\bar{x}_{Q'+n} - \bar{x}_n) \\ &= \alpha + \beta\bar{x}_n + \beta\,(\bar{x}_N - \bar{x}_n) \\ &= \alpha + \beta\bar{x}_N \\ &= \bar{y}_N \end{aligned} \tag{123}$$

To obtain its variance, we write

$$\begin{aligned} V(\bar{y}'_{ds}) &= E\{\bar{y}_n - \alpha - \beta\bar{x}_N + b\,(\bar{x}_{Q'+n} - \bar{x}_n)\}^2 \\ &= E\{\bar{y}_n - \alpha - \beta\bar{x}_n + (b-\beta)\,(\bar{x}_{Q'+n} - \bar{x}_n) \\ &\qquad\qquad + \beta\,(\bar{x}_{Q'+n} - \bar{x}_N)\}^2 \\ &= E\{\bar{y}_n - \alpha - \beta\bar{x}_n\}^2 + E\{(b-\beta)^2\,(\bar{x}_{Q'+n} - \bar{x}_n)^2\} \\ &\qquad\qquad + \beta^2 E\,(\bar{x}_{Q'+n} - \bar{x}_N)^2 \\ &= \frac{\gamma}{n} + \gamma E\left\{\frac{(\bar{x}_{Q'+n} - \bar{x}_n)^2}{\sum\limits_{i=1}^{k} n_i\,(x_i - \bar{x}_n)^2}\right\} + \beta^2 E\,(\bar{x}_{Q'+n} - \bar{x}_N)^2 \end{aligned} \tag{124}$$

If we assume a normal distribution for the x's, we have, from (43),

$$\begin{aligned} & E\left\{\frac{(\bar{x}_{Q'+n} - \bar{x}_n)^2}{\sum\limits_{i=1}^{k} n_i\,(x_i - \bar{x}_n)^2}\right\} \\ &= E\{(\bar{x}_{Q'+n} - \bar{x}_n)^2\}\cdot E\left\{\frac{1}{\sum\limits_{i=1}^{k} n_i(x_i - \bar{x}_n)^2}\right\} \\ &= \left(\frac{1}{n} - \frac{1}{Q'+n}\right)\frac{1}{n-3} \end{aligned} \tag{125}$$

By a method similar to that adopted in Section 5.4, it can, however, be shown that (125) will be true in general even without assuming normality, provided terms of higher order in $1/n$ are ignored. Hence

$$V(\bar{y}'_{ds}) \simeq \frac{\gamma}{n}\left\{1 + \frac{Q'}{Q'+n}\cdot\frac{1}{n-3}\right\} + \frac{\rho^2 S_y^2}{Q'+n} \tag{126}$$

which is seen to be smaller than (121). The expression was first given by Cochran (*see* Jessen, R. J., 1942).

The estimation of the variance of either estimate presents no difficulty. Thus, to estimate the variance of $\bar{y}_{ds}$, we note that

$$s^2_{x|Q'} = \frac{\sum\limits^{Q'} (x_i - \bar{x}_{Q'})^2}{Q'-1}$$

is an unbiased estimate of $\mathbf{S}^2{}_{x|N'}$, from (112) b^2 is an unbiased estimate of

$$\beta^2 + \frac{\gamma}{\sum\limits_{i=1}^{k} n_i (x_i - \bar{x}_n)^2}$$

from (115)

$$\text{Est. } (\bar{x}_N - \bar{x}_n)^2 = \frac{N'^2}{N^2} \left\{(\bar{x}_{Q'} - \bar{x}_n)^2 - \left(\frac{1}{Q'} - \frac{1}{N'}\right) s^2{}_{x|Q'}\right\}$$

and the estimate of γ is given by (36) in Section 5.3. On making the substitutions in (117), we get the desired estimate.

Similarly, we have from (126),

$$\text{Est. } V(\bar{y}'_{ds}) = \frac{\mathbf{Q}}{n(n-2)} \left\{1 + \frac{Q'}{Q'+n} \cdot \frac{1}{n-3}\right\} + \frac{1}{Q'+n} \left\{s_y{}^2 - \frac{\mathbf{Q}}{n-2}\right\} \qquad (127)$$

5.11* Successive Sampling

In using the method of regression we have so far assumed that ancillary information is available for all the units in the sample for which the character under study is observed. This is not, however, always the case and the formulæ given above need to be extended to make use of the information contained in additional units recording only the character under study. This is the case, for instance, when the same variate is observed on two successive occasions in time, there being in the units observed some which are common to both occasions and some which are exclusive to each occasion. The observations on the earlier occasion are used as ancillary information to improve the estimate of the population value on the second occasion. We shall assume the population to be large.

Let the character be observed on $Q' + n_1$ units on the first occasion and $n_1 + n_2$ units on the second occasion, of which n_1 are common to the first occasion. We may, as in (122), form a regression estimate of the mean on the second occasion based on the n_1 units, *viz.*,

$$\bar{z} = \bar{y}_{n_1} + b\,(\bar{x}_{Q'+n_1} - \bar{x}_{n_1}) \qquad (128)$$

where y denotes observations on the second occasion and x those on the first occasion. Another independent estimate of the population mean on the second occasion is available on the basis of n_2 units observed on the second occasion only, *viz.*, $\bar{y}_{n_2}$. Any unbiased linear combination of the two estimates can be written as:

$$\bar{y} = (1 - \psi)\,\bar{z} + \psi\bar{y}_{n_2} \tag{129}$$

where ψ is any positive number less than 1. To obtain the best overall estimate, we may choose ψ so as to minimize the variance of (129), *i.e.*, minimize the expression

$$(1 - \psi)^2\, V(\bar{z}) + \psi^2\, V(\bar{y}_{n_2})$$

We have easily

$$\psi V(\bar{y}_{n_2}) = (1 - \psi)\, V(\bar{z}) \tag{130}$$

Thus

$$\psi = \frac{V(\bar{z})}{V(\bar{y}_{n_2}) + V(\bar{z})} \tag{131}$$

where, from (126),

$$V(\bar{z}) \cong \frac{\mathbf{S}_y^2}{n_1}(1 - \rho^2)\cdot\left[1 + \frac{Q'}{Q' + n_1}\cdot\frac{1}{n_1 - 3}\right] + \frac{\rho^2\mathbf{S}_y^2}{Q' + n_1}$$

and

$$V(\bar{y}_{n_2}) = \frac{\mathbf{S}_y^2}{n_2}$$

When information is available for more than one previous occasion, the estimate on the current occasion can be worked out through a recurrence formula based on precisely the same arguments as given above for two occasions. Denote by $\bar{y}_h$ the best available estimate of the population mean on the h-th occasion, using observations up to the h-th occasion. We write

$$\bar{y}_h = (1 - \psi_h)\,\bar{z}_h + \psi_h\bar{y}_{n''_h} \tag{132}$$

where

$$\bar{z}_h \quad = \bar{y}_{n'_h} + b_{h,\,h-1}\,(\bar{y}_{h-1} - \bar{x}_{n'_h})$$

$\bar{y}_{n'_h}$ = mean on the h-th occasion of n'_h units common to the $(h - 1)$-th occasion

$\bar{x}_{n'_h}$ = mean of the same n'_h units on the preceding occasion

$\bar{y}_{n''_h}$ = mean on the h-th occasion of n''_h units not common with the $(h-1)$-th occasion

$b_{h,\,h-1}$ = regression coefficient of values of the h-th occasion on values of the $(h-1)$-th occasion based on the n'_h common units,

$$= \frac{\sum^{n'_h} y\,(x - \bar{x}_{n'_h})}{\sum^{n'_h} (x - \bar{x}_{n'_h})^2}$$

As before it will be seen that $\bar{z}_h$ is the regression estimate of the population mean μ_h on the h-th occasion, using observations on the $(h-1)$-th occasion as ancillary information. Another independent estimate is provided by $\bar{y}_{n''_h}$ and we choose ψ_h so as to minimize the variance of the above combination. For presentation of the exact theory it will be assumed that units in the sample are common only between consecutive occasions, but the results will, in practice, be still true if it can be assumed that observations more than one occasion apart do not provide any additional information. Minimizing the variance with respect to ψ_h, we have

$$\psi_h V(\bar{y}_{n''_h}) = (1 - \psi_h)\, V(\bar{z}_h) \tag{133}$$

Now, using (124) and putting $\gamma = S_h{}^2\,(1 - \rho^2{}_{h,h-1})$, we write

$$V(\bar{z}_h) = \frac{\gamma}{n'_h} + \gamma\, E\left\{\frac{(\bar{y}_{h-1} - \bar{x}_{n'_h})^2}{\sum^{n'_h}(x - \bar{x}_{n'_h})^2}\right\} + \beta^2{}_{h,h-1} E\,(\bar{y}_{h-1} - \mu_{h-1})^2 \tag{134}$$

Assuming independence of $\sum^{n'_h}(x - \bar{x}_{n'_h})^2$ and $(\bar{y}_{h-1} - \bar{x}_{n'_h})^2$, which will necessarily be true in case of normality, we have

$$V(\bar{z}_h) = \frac{\mathrm{S}_h{}^2\,(1 - \rho^2{}_{h,\,h-1})}{n'_h}$$

$$+ \frac{\mathrm{S}_h{}^2\,(1 - \rho^2{}_{h,\,h-1})}{\mathrm{S}^2{}_{h-1}\,(n'_h - 3)}\, E\,(\bar{y}_{h-1} - \bar{x}_{n'_h})^2 + \beta^2{}_{h,h-1}\, V(\bar{y}_{h-1}) \tag{135}$$

But

$$E\,(\bar{y}_{h-1} - \bar{x}_{n'_h})^2$$

$$= V\,(\bar{y}_{h-1}) + \frac{\mathbf{S}^2_{h-1}}{n'_h} - 2\,\text{Cov}\,(\bar{y}_{h-1}, \bar{x}_{n'_h})$$

$$\cong \frac{\mathbf{S}^2_{h-1}}{n'_h} - \psi_{h-1}\,\frac{\mathbf{S}^2_{h-1}}{n''_{h-1}} \tag{136}$$

since

$$\text{Cov}\,\,(\bar{y}_{h-1}, \bar{x}_{n'_h}) = V\,(\bar{y}_{h-1}) = \psi_{h-1}\,\frac{\mathbf{S}^2_{h-1}}{n''_{h-1}} \tag{137}$$

The validity of (137) follows from the well-known result that the correlation between any unbiased estimate and an efficient estimate tends to the square root of the ratio of their variances. The approximation involved will affect only terms of order $(1/n'_h)^3$ in the expression for $V\,(\bar{z}_h)$. The result is, however, exact if units are common only between two consecutive occasions, for, we have in that case

$$\text{Cov}\,(\bar{y}_{h-1}, \bar{x}_{n'_h})$$

$$= \text{Cov}\,[\{(1 - \psi_{h-1})\,\bar{z}_{h-1} + \psi_{h-1}\bar{y}_{n''_{h-1}}\}, \bar{x}_{n'_h}]$$

$$= \psi_{h-1}\,\text{Cov}\,(\bar{y}_{n''_{h-1}}, \bar{x}_{n'_h}), \text{ since } \text{Cov}\,(\bar{z}_{h-1}, \bar{x}_{n'_h}) = 0$$

$$= \psi_{h-1}\,E\,\{(\bar{y}_{n''_{h-1}} - \mu_{h-1})\,(\bar{x}_{n'_h} - \bar{y}_{n''_{h-1}} + \bar{y}_{n''_{h-1}} - \mu_{h-1})\}$$

$$= \psi_{h-1} V\,(\bar{y}_{n''_{h-1}})$$

since

$$E\,(\bar{y}_{n''_{h-1}} - \mu_{h-1})\,(\bar{x}_{n'_h} - \bar{y}_{n''_{h-1}}) = 0$$

Thus

$$\text{Cov}\,\,(\bar{y}_{h-1},\,\bar{x}_{n'_h}) = \psi_{h-1}\,\frac{\mathbf{S}^2_{h-1}}{n''_{h-1}} \tag{138 }$$

Hence we have, from (133), (135), (136) and (137),

$$\psi_h\,\frac{\mathbf{S}_h^{\,2}}{n''_h} = (1 - \psi_h)\left[\frac{\mathbf{S}_h^{\,2}\,(1 - \rho^2_{h,\,h-1})}{n'_h} + \frac{\mathbf{S}_h^{\,2}\,(1 - \rho^2_{h,\,h-1})}{n'_h - 3}\right.$$

$$\left.\times \left\{\frac{1}{n'_h} - \psi_{h-1}\,\frac{1}{n''_{h-1}}\right\} + \rho^2_{h,\,h-1}\,\mathbf{S}_h^{\,2}\,\frac{\psi_{h-1}}{n''_{h-1}}\right] \tag{139}$$

Writing

$$\rho'^2_{h,\,h-1} \equiv \rho^2_{h,\,h-1} - \frac{1-\rho^2_{h,\,h-1}}{n'_h - 3} \tag{140}$$

equation (139) can be expressed as

$$(1-\psi_h)\left\{\frac{1-\rho'^2_{h,\,h-1}}{n'_h} + \frac{\rho'^2_{h,\,h-1}\psi_{h-1}}{n''_{h-1}}\right\} = \psi_h \frac{1}{n''_h} \tag{141}$$

which provides the recurrence formula for the evaluation of ψ_h (remembering that $\psi_1 = 1$) and hence of the estimate (132). In practice, the correlation coefficients $\rho_{h,\,h-1}$ may have to be evaluated from the sample. The recurrence formula (141) is due to Narain (1953). The general theory has also been investigated by Yates (1949), Patterson (1950) and Tikkiwal (1951), but the regression coefficients $\beta_{h,\,h-1}$ have been assumed known in their approach.

Finally, we shall consider the question of replacement. What fraction of the sample should be replaced on each occasion in order that the estimate on the current occasion may have the maximum precision? We shall first consider the case of two occasions and assume that: (*a*) the total size of sample $n_1 + n_2$ is fixed on each occasion, say equal to M, and (*b*) n_1 is sufficiently large so as to neglect terms in $1/n_1{}^2$ in the expression for $V(\bar{z})$. Substituting for ψ from (131) and for $V(\bar{z})$ and $V(\bar{y}_{n_2})$ in the expression for the variance of $\bar{y}$, we obtain

$$\begin{aligned}
V(\bar{y}) &= (1-\psi)^2\, V(\bar{z}) + \psi^2 V(\bar{y}_{n_2}) \\
&= \left(1 - \frac{V(\bar{z})}{V(\bar{y}_{n_2}) + V(\bar{z})}\right)^2 V(\bar{z}) \\
&\qquad\qquad + \left(\frac{V(\bar{z})}{V(\bar{y}_{n_2}) + V(\bar{z})}\right)^2 V(\bar{y}_{n_2}) \\
&= \frac{V(\bar{y}_{n_2})\, V(\bar{z})}{V(\bar{y}_{n_2}) + V(\bar{z})} \\
&= \frac{\mathrm{S}_y{}^2\left(1 - \frac{n_2}{M}\rho^2\right)}{M\left(1 - \frac{n_2{}^2}{M^2}\rho^2\right)}
\end{aligned} \tag{142}$$

Clearly, the optimum value of the fraction to be replaced is obtained by minimizing $V(\bar{y})$ with respect to n_2/M. Differentiating with respect to n_2/M and equating to zero and noting that $(n_1 + n_2)/M = 1$, we obtain

$$\frac{n_2}{M} = \frac{1}{1 + \sqrt{1 - \rho^2}} \tag{143}$$

It will be noticed that the fraction to be replaced depends upon the value of ρ. The larger the value of ρ, the larger is clearly the fraction to be replaced. n_2/M attains the minimum value of $\frac{1}{2}$ when $\rho = 0$, showing that the fraction to be replaced should always exceed $\frac{1}{2}$ provided, of course, cost and practical considerations warrant such replacement. For moderately high values of ρ like $\cdot 5$ to $\cdot 7$, the optimum value of the fraction to be replaced works out to about 3/5 of the size of the total sample.

Now, in general, on the h-th occasion, the variance of $\bar{y}_h$ will be seen to be given by

$$V(\bar{y}_h) = (1 - \psi_h)^2 \, V(\bar{z}_h) + \psi_h{}^2 \, V(\bar{y}_{n''_h})$$

$$= (1 - \psi_h)^2 \, \frac{\psi_h}{1 - \psi_h} \, V(\bar{y}_{n''_h}) + \psi_h{}^2 \, V(\bar{y}_{n''_h})$$

by virtue of (133), *i.e.*,

$$= \psi_h \, \frac{\mathrm{S}_h{}^2}{n''_h} \tag{144}$$

Let h be sufficiently large so as to justify the use of the limiting value of ψ_h obtained by putting $\psi_h = \psi_{h-1}$ in (141), and further let $n'_h = n'$ and $n''_h = n''$ for all h. Differentiating (144) with respect to n''/M, where $M = n' + n''$, and using the limiting value of ψ, *viz.*,

$$\psi = \frac{-(1 - \rho^2) + \sqrt{(1 - \rho^2)\left\{1 - \rho^2 + 4\rho^2 \dfrac{n'n''}{M^2}\right\}}}{2\, \dfrac{n'}{M}\, \rho^2}$$

we obtain

$$\frac{n''}{M} = \frac{1}{2} \tag{145}$$

thus showing that the replacement fraction to be used after a sufficiently large number of occasions is $\frac{1}{2}$ (Tikkiwal, 1951).

REFERENCES

1. Neyman, J. (1934) .. "On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection," *Jour. Roy. Statist. Soc.*, **97**, 558–606.

2. —— and David, F. N. (1938) "Extension of the Markoff Theorem on Least Squares," *Statist. Res. Mem.*, **2**, 105–16.

3. Hasel, A. A. (1942) .. "Estimation of Volume in Timber Stands by Strip Sampling," *Ann. Math. Statist.*, **13**, 179–206.

4. Sukhatme, P. V. (1944) .. "Moments and Product Moments of Moment-Statistics for Samples of the Finite and Infinite Populations," *Sankhya*, **6**, 363–82.

5. Cochran, W. G. (1942) .. "Sampling Theory when the Sampling Units are of Unequal Sizes," *Jour. Amer. Statist. Assoc.*, **37**, 199–212.

6. Yates, F. (1949) .. *Sampling Methods for Censuses and Surveys*, Charles Griffin & Co., Ltd., London.

7. Patterson, H. D. (1950) .. "Sampling on Successive Occasions with Partial Replacement of Units," *Jour. Roy. Statist. Soc.*, *Series B*, **12**, 241–55.

8. Narain, R. D. (1953) .. "On the Recurrence Formula in Sampling on Successive Occasions," *Jour. Ind. Soc. Agr. Statist.*, **5**, No. 1. *In Press.*

9. Tikkiwal, B. D. (1951) .. "Theory of Successive Sampling," Unpublished Thesis for Diploma, I.C.A.R., New Delhi.

10. Sukhatme, P. V. (1953) .. "The Variance of the Regression Estimate in Double Sampling from Finite Populations," *Metron*, **17**, Nos. 1–2. *In Press.*

CHAPTER VI

CHOICE OF SAMPLING UNIT

A. EQUAL CLUSTERS

6a.1 Cluster Sampling

A sampling procedure, as pointed out in Section 1.4, presupposes the division of the population into a finite number of distinct and identifiable units called the sampling units. Thus a population of fields under wheat in a given region might be regarded as composed of fields or groups of fields on the same holdings, villages, or other suitable segments. A human population might similarly be regarded as composed of individual persons, families, or groups of persons residing in houses and villages. The smallest units into which the population can be divided are called the *elements* of the population, and groups of elements the *clusters*. When the sampling unit is a cluster, the procedure of sampling is called *cluster sampling*. When the entire area containing the population under study is subdivided into smaller areas and each element in the population is associated with one and only one such small area, the procedure is alternatively called *area sampling*.

For many types of population a list of elements is not available and the use of an element as the sampling unit is therefore not feasible. The method of cluster or area sampling is available in such cases. Thus, in a city a list of all the houses is readily available, but that of persons is rarely so. Again, lists of fields are not available, but those of villages are. Cluster sampling is, therefore, widely practised in sample surveys.

The size of the cluster to be employed in sample surveys therefore requires consideration. In general, the smaller the cluster, the more accurate will usually be the estimate of the population character for a given number of elements in the sample. Thus, a sample of holdings independently and randomly selected is likely to be scattered over the entire area under the crop, and thereby provides a better cross-section of the population than an equivalent sample, *i.e.*, a sample of the same number of holdings, clustered together in a few villages. On the other hand, it will cost more to survey a widely scattered sample of holdings than to survey an equivalent sample of clusters of holdings, since the additional cost of surveying a neighbouring holding is small as compared to the cost of locating a second independent holding and surveying it. The optimum cluster is one which gives an estimate of the character under study with the smallest standard error for a given proportion of the population sampled, or more generally, for a given cost. In this chapter, we will give the relevant theory which can provide guidance in the choice of a sampling unit in a sample survey.

6*a*.2 Efficiency of Cluster Sampling

We shall first consider the case of equal clusters and suppose that the population is composed of N clusters of M elements each, and that a sample of n clusters is drawn from it by the method of simple random sampling.

Let

y_{ij} denote the value of the character for the j-th element, ($j = 1, 2, \ldots, M$) in the i-th cluster, ($i = 1, 2, \ldots, N$);

$\bar{y}_{i.}$ the mean per element of the i-th cluster, given by

$$\bar{y}_{i.} = \frac{1}{M} \sum_{j=1}^{M} y_{ij} \tag{1}$$

$\bar{y}_{N.}$ the mean of cluster means in the population, defined by

$$\bar{y}_{N.} = \frac{1}{N} \sum_{i=1}^{N} \bar{y}_{i.} \tag{2}$$

$\bar{y}_{..}$ the mean per element in the population, defined by

$$\bar{y}_{..} = \frac{1}{NM} \sum_{i=1}^{N} \sum_{j=1}^{M} y_{ij} \tag{3}$$

and

$\bar{y}_{n.}$ the mean of cluster means in a simple random sample of n clusters, given by

$$\bar{y}_{n.} = \frac{1}{n}\sum^{n} \bar{y}_{i.} \tag{4}$$

= the mean per element in the sample.

Clearly, $\bar{y}_{n.}$ is an unbiased estimate of $\bar{y}_{..}$, and its variance is

$$V(\bar{y}_{n.}) = \frac{N-n}{Nn}\,\mathrm{S}_b^2 \tag{5}$$

where

$$\mathrm{S}_b^2 = \frac{1}{N-1}\sum_{i=1}^{N} (\bar{y}_{i.} - \bar{y}_{N.})^2 \tag{6}$$

= the mean square between cluster means in the population.

If an equivalent sample of nM elements were independently selected from the population, the variance of the mean per element would be

$$V(\bar{y}_{nM}) = \frac{NM-nM}{NM}\cdot\frac{\mathrm{S}^2}{nM} \tag{7}$$

where

$$\mathrm{S}^2 = \frac{1}{NM-1}\sum_{i=1}^{N}\sum_{j=1}^{M} (y_{ij} - \bar{y}_{N.})^2 \tag{8}$$

= the mean square between elements in the population.

Now the relative efficiency E of any two estimates based on samples of equal size is defined as the ratio of the inverse of their variances. It follows that the efficiency of a cluster as the sampling unit compared with that of an element is given by

$$E = \frac{V(\bar{y}_{nM})}{V(\bar{y}_{n.})} \tag{9}$$

$$= \frac{\dfrac{NM-nM}{NM} \cdot \dfrac{S^2}{nM}}{\dfrac{N-n}{Nn} \cdot S_b^2}$$

$$= \frac{S^2}{MS_b^2} \tag{10}$$

If we set up an analysis of variance for elements in the population, as shown in Table 6.1, this efficiency will be seen to be equal to the ratio of the overall mean square between elements to that between clusters in the population.

TABLE 6.1

Analysis of Variance

Source of Variation		Degrees of Freedom	Mean Square
Between clusters	..	$N-1$	$\frac{M}{N-1}\sum_{i=1}^{N}(\bar{y}_{i.}-\bar{y}_{N.})^2 = MS_b^2$
Within clusters	..	$N(M-1)$	$\frac{1}{N(M-1)}\sum_{i=1}^{N}\sum_{j=1}^{M}(y_{ij}-\bar{y}_{i.})^2 = \bar{S}_w^2$
Total population	..	$NM-1$	$\frac{1}{NM-1}\sum_{i=1}^{N}\sum_{j=1}^{M}(y_{ij}-\bar{y}_{N.})^2 = S^2$

Example 6.1

To show how efficiency changes with the size of a cluster, we give a numerical example from data relating to the use of clusters of different sizes in estimating the area under wheat. Table 6.2 gives values of the mean square between survey numbers in a village (S^2) and the mean square between clusters (MS_b^2) pooled over 11 villages in the Meerut District (India).

The clusters were formed by grouping consecutive survey numbers in a village. The character studied was the area under wheat. The mean squares are given separately for clusters of size 2, 4, 8 and 16 survey numbers. The last row of the table gives the values of the efficiency obtained by dividing the mean square between survey numbers by that between clusters within villages.

TABLE 6.2

Efficiency of Clusters of Size 2, 4, 8 and 16 Survey Numbers

Mean Square (Acres)^2	Size of Cluster			
	2	4	8	16
Between clusters within villages ..	138·6	180·7	245·1	333·9
Between survey numbers within villages	108·3	108·3	108·3	108·3
Efficiency	0·78	0·60	0·44	0·32

It will be seen that the efficiency decreases rather rapidly with the increase in the size of the cluster, clusters of 2 being only about four-fifths as efficient as individual survey numbers, those of 4 about three-fifths, while those of 16 are only one-third as efficient as individual survey numbers. In other words, the sample of clusters of size 16 will have to be three times as large as the sample of individual survey numbers in order to give an estimate of equal precision.

If clusters are random samples of M from a population of NM elements, and consequently composed of elements which are not more alike than those of other clusters, then the mean squares between and within clusters will behave as random variables and their expected values will each be of the same order.

For,

E (Mean square between clusters)

$$= E\left\{\frac{1}{N-1}\sum_{i=1}^{N} M(\bar{y}_{i.}-\bar{y}_{N.})^2\right\}$$

$$= \frac{M}{N-1}\left\{\sum_{i=1}^{N} E(\bar{y}_{i.}^2) - N\bar{y}_{N.}^2\right\}$$

Now substituting from (35) in Section 2*a*.5, we have

E (Mean square between clusters)

$$= \frac{M}{N-1}\left\{\sum_{i=1}^{N}\left(\bar{y}_{N.}^{2} + \frac{NM-M}{NM}\cdot\frac{\mathbf{S}^2}{M}\right) - N\bar{y}_{N.}^{2}\right\}$$

$$= \mathbf{S}^2 \tag{11}$$

Similarly

E (Mean square within clusters)

$$= E\left\{\frac{1}{N(M-1)}\sum_{i=1}^{N}\sum_{j=1}^{M}(y_{ij}-\bar{y}_{i.})^2\right\}$$

$$= \frac{1}{N(M-1)}\left\{\sum_{i=1}^{N}\sum_{j=1}^{M}E(y_{ij}^{2}) - M\sum_{i=1}^{N}E(\bar{y}_{i.}^{2})\right\}$$

$$= \frac{1}{N(M-1)}\left\{NM\left(\bar{y}_{N.}^{2} + \frac{NM-1}{NM}\mathbf{S}^2\right) - MN\left(\bar{y}_{N.}^{2} + \frac{NM-M}{NM}\cdot\frac{\mathbf{S}^2}{M}\right)\right\}$$

$$= \mathbf{S}^2 \tag{12}$$

It follows that if clusters are random samples of the population, they will, on the average, be as efficient as the individual elements themselves.

6*a*.3 Efficiency of Cluster Sampling in Terms of Intra-Class Correlation

In practice, a cluster cannot be regarded as comprised of a random sample of elements of the population. Usually, elements of the same cluster will resemble each other more than those belonging to different clusters. Thus, the variation in yield between different portions of a field will tend to be less than that between different fields. Consequently, the variance of an estimate based on cluster sampling will ordinarily exceed that based on an equivalent sample of elements selected independently. The manner in which the variance of the estimate increases with the size of the cluster can best be elucidated with the help of the concept of intra-class correlation between elements of a cluster.

Let ρ denote the intra-class correlation defined by

$$\rho = \frac{E\,\{(y_{ij}-\bar{y}_{N.})\,(y_{ik}-\bar{y}_{N.})\}}{E\,(y_{ij}-\bar{y}_{N.})^2} \tag{13}$$

The numerator in (13) can be written as

$$\begin{aligned} &E\,\{(y_{ij}-\bar{y}_{i.}+\bar{y}_{i.}-\bar{y}_{N.})\,(y_{ik}-\bar{y}_{i.}+\bar{y}_{i.}-\bar{y}_{N.})\} \\ &\quad = E\,\{(y_{ij}-\bar{y}_{i.})\,(y_{ik}-\bar{y}_{i.})+(y_{ik}-\bar{y}_{i.})\,(\bar{y}_{i.}-\bar{y}_{N.}) \\ &\qquad\qquad +(y_{ij}-\bar{y}_{i.})\,(\bar{y}_{i.}-\bar{y}_{N.})+(\bar{y}_{i.}-\bar{y}_{N.})^2\} \\ &\quad = E\,\{(y_{ij}-\bar{y}_{i.})\,(y_{ik}-\bar{y}_{i.})\}+E\,(\bar{y}_{i.}-\bar{y}_{N.})^2 \end{aligned} \tag{14}$$

since the expected value of the two middle terms is clearly zero. To evaluate the first term of (14), we first work out the expectation for a given i. We have

$$\begin{aligned} &E\,\{(y_{ij}-\bar{y}_{i.})\,(y_{ik}-\bar{y}_{i.})\mid i\} \\ &\quad = \frac{1}{M\,(M-1)}\sum_{j\neq k=1}^{M}(y_{ij}-\bar{y}_{i.})\,(y_{ik}-\bar{y}_{i.}) \\ &\quad = \frac{1}{M\,(M-1)}\left[\left\{\sum_{j=1}^{M}(y_{ij}-\bar{y}_{i.})\right\}^2-\sum_{j=1}^{M}(y_{ij}-\bar{y}_{i.})^2\right] \\ &\quad = \frac{1}{M\,(M-1)}\left[0-(M-1)\,\mathbf{S}_i^2\right] \\ &\quad = -\frac{\mathbf{S}_i^2}{M} \end{aligned} \tag{15}$$

Next, taking the expectations for varying i, we obtain

$$\begin{aligned} E\,\{(y_{ij}-\bar{y}_{i.})\,(y_{ik}-\bar{y}_{i.})\} &= -\frac{1}{N}\sum_{i=1}^{N}\frac{\mathbf{S}_i^2}{M} \\ &= -\frac{\bar{\mathbf{S}}_w^2}{M} \end{aligned} \tag{16}$$

where

$$\bar{S}_w^2 = \frac{1}{N} \sum_{i=1}^{N} S_i^2.$$

The values of the second term in (14) and that of the denominator in (13) are known, by definition, to be $(N-1)\,S_b^2/N$ and $(NM-1)\,S^2/NM$ respectively. We thus have

$$\rho = \frac{\frac{N-1}{N} S_b^2 - \frac{\bar{S}_w^2}{M}}{\frac{NM-1}{NM} S^2} \tag{17}$$

When clusters are randomly formed, the expected value of $\bar{S}_w^2$ and MS_b^2 will each be equal to S^2, and

$$\rho = -\frac{1}{NM-1} \tag{18}$$

Now, by definition, S^2, S_b^2, and $\bar{S}_w^2$ are related to each other by the identity

$$(NM-1)\,S^2 \equiv (N-1)\,MS_b^2 + N\,(M-1)\,\bar{S}_w^2 \tag{19}$$

Hence, eliminating first $\bar{S}_w^2$ from (17) and (19), we have

$$S_b^2 = \frac{NM-1}{M\,(N-1)} \cdot \frac{S^2}{M} \left\{1 + (M-1)\,\rho\right\} \tag{20}$$

and next eliminating S_b^2 from the same equations, we get

$$\bar{S}_w^2 = \frac{NM-1}{NM}\, S^2\,(1-\rho) \tag{21}$$

The variance of the mean of n cluster means is now obtained by substituting from (20) in (5). We have

$$V\,(\bar{y}_{n.}) = \frac{N-n}{N} \cdot \frac{S_b^2}{n}$$

$$= \frac{N-n}{N} \cdot \frac{NM-1}{M\,(N-1)} \cdot \frac{S^2}{nM} \left\{1 + (M-1)\,\rho\right\} \tag{22}$$

and from (10) the relative efficiency is given by

$$E = \frac{M(N-1)}{NM-1} \cdot \frac{1}{\{1+(M-1)\rho\}} \tag{23}$$

For N large, the formulæ (17) and (19) to (23) can be approximated by the simpler expressions

$$\rho \cong \frac{\mathbf{S}_b^2 - \frac{\bar{\mathbf{S}}_w^2}{M}}{\mathbf{S}^2} \tag{24}$$

$$\mathbf{S}^2 \cong \mathbf{S}_b^2 + \frac{M-1}{M} \bar{\mathbf{S}}_w^2 \tag{25}$$

$$\mathbf{S}_b^2 \cong \frac{\mathbf{S}^2}{M} \left\{1 + (M-1)\rho\right\} \tag{26}$$

$$\bar{\mathbf{S}}_w^2 \cong \mathbf{S}^2 (1-\rho) \tag{27}$$

$$V(\bar{y}_{n.}) \cong \frac{N-n}{N} \cdot \frac{\mathbf{S}^2}{nM} \left\{1 + (M-1)\rho\right\} \tag{28}$$

and

$$E \cong \frac{1}{1+(M-1)\rho} \tag{29}$$

Formula (22) for the variance of the mean of n cluster means can be expressed more simply by introducing an alternative notation. Let the variance of the mean of a single cluster be denoted by σ_b^2, so that

$$\sigma_b^2 = \frac{N-1}{N} \mathbf{S}_b^2 \tag{30}$$

that of a single element in a specified cluster by $\bar{\sigma}_w^2$, so that

$$\bar{\sigma}_w^2 = \frac{M-1}{M} \bar{\mathbf{S}}_w^2 \tag{31}$$

and that of a single element chosen from the population by σ^2, so that

$$\sigma^2 = \frac{NM-1}{NM} \mathbf{S}^2 \tag{32}$$

On substituting for $\mathbf{S}_b^2$; $\bar{\mathbf{S}}_w^2$ and $\mathbf{S}^2$ from (30), (31) and (32) in (17), (19), (20), (21) and (22), we have

$$\rho = \frac{1}{\sigma^2}\left(\sigma_b^2 - \frac{\bar{\sigma}_w^2}{M-1}\right) \tag{33}$$

$$\sigma^2 = \sigma_b^2 + \bar{\sigma}_w^2 \tag{34}$$

$$\sigma_b^2 = \frac{\sigma^2}{M}\left\{1 + (M-1)\,\rho\right\} \tag{35}$$

$$\bar{\sigma}_w^2 = \frac{M-1}{M}\cdot\sigma^2\cdot(1-\rho) \tag{36}$$

and

$$V(\bar{y}_{n.}) = \frac{N-n}{N-1}\cdot\frac{\sigma^2}{nM}\left\{1 + (M-1)\,\rho\right\} \tag{37}$$

The formula for the variance of the mean of n cluster means in this form was first developed by Hansen and Hurwitz (1942). It is made up of three factors. The first factor is the finite multiplier $(N-n)/(N-1)$. The second factor is the variance of the mean based on nM elements selected with replacement at random. The third factor measures the contribution to the variance of cluster sampling. If $M = 1$, this factor is unity and we are left with the finite multiplier and the sampling variance of the mean per element based on nM elements selected independently at random. If M is greater than 1, $(M-1)\,\rho$ will, therefore, measure the relative change in the sampling variance brought about by sampling clusters instead of elements. In practice, ρ is usually positive and decreases as M increases, but the rate of decrease is small relative to the rate of increase in M, so that ordinarily increase in the size of a cluster leads to substantial increase in the sampling variance of the sample estimate.

The point can be well illustrated with the help of data given in Table 6.3. Table 6.3 gives values of ρ and $(M-1)\,\rho$ calculated from the data in Table 6.2. It will be seen that ρ decreases as M increases as expected, but the rate of decrease in ρ is slow as compared to the rate of increase in M. For clusters of size 16, the relative increase in sampling variance is seen to be a little over 200%.

TABLE 6.3

Relative Change in Variance with Increase in Size of Cluster

Size of Cluster (M)	2	4	8	16
ρ	0·28	0·22	0·18	0·14
$(M-1)\,\rho$	0·28	0·66	1·26	2·10

Although ρ is usually positive, there are situations where it can also be negative, as, for example, in the problem of estimating the proportion of males in the population using the household as a sampling unit. The intra-class correlation between sexes of different members of a household is clearly negative. Consider, for instance, households of size 4, consisting of a husband, a wife and two children. The households can be classified into three classes: those in which both children are males, those in which one child is a male and the other is a female, and finally, those in which both children are females. On the assumption that the proportion of male children is one-half, we would expect both children to be males in 25% of the households, one child to be male in 50% of the households, and no child to be male in the remaining 25%. These relative frequencies p_i ($i = 1, 2, 3$), together with the values of the proportion of males in the different classes *viz.*, $\bar{y}_{i.}$, are presented in Table 6.4. In this

TABLE 6.4

Relative Frequencies and Proportions of Males in Households of Size 4

Class	Description of Household	Frequency p_i	Proportion of Males $\bar{y}_{i.}$	$\dfrac{\sum\limits_{j=1}^{4}(y_{ij}-\frac{1}{2})^2}{4}$	$\dfrac{\sum\limits_{j=1}^{4}(y_{ij}-\bar{y}_{i.})^2}{4}$	$(\bar{y}_{i.}-\frac{1}{2})^2$
1	2 male children	$\frac{1}{4}$	$\frac{3}{4}$	$\frac{1}{4}$	$\frac{3}{16}$	$\frac{1}{16}$
2	1 male child	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$	0
3	No male children	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{3}{16}$	$\frac{1}{16}$
Total population			$\frac{1}{2}$	$\frac{1}{4}$	$\frac{7}{32}$	$\frac{1}{32}$

table $y_{ij} = 1$, if the j-th member in the i-th class is a male, and $y_{ij} = 0$, if it is a female.

Since the population under consideration is an infinite population, we write

$$S^2 = \sigma^2 = E\{y_{ij} - E(y_{ij})\}^2$$

$$= \sum_{i=1}^{3} p_i \frac{\sum_{j=1}^{4} (y_{ij} - \frac{1}{2})^2}{4}$$

$$= \frac{1}{4} \tag{38}$$

Also

$$\bar{S}_w^2 = \bar{\sigma}_w^2 = \sum_{i=1}^{3} p_i \frac{\sum_{j=1}^{4} (y_{ij} - \bar{y}_{i.})^2}{4}$$

$$= \frac{7}{32} \tag{39}$$

and

$$S_b^2 = \sigma_b^2 = \sum_{i=1}^{3} p_i (\bar{y}_{i.} - \tfrac{1}{2})^2$$

$$= \frac{1}{32} \tag{40}$$

The values of $\frac{1}{4} \sum_{j=1}^{4} (y_{ij} - \frac{1}{2})^2$, $\frac{1}{4} \sum_{j=1}^{4} (y_{ij} - \bar{y}_{i.})^2$ and $(\bar{y}_{i.} - \frac{1}{2})^2$ for the different classes, and those for the whole population representing σ^2, $\bar{\sigma}_w^2$ and σ_b^2 are also shown in Table 6.4. On substituting in (33), we obtain

$$\rho = \frac{(\frac{1}{32}) - (\frac{1}{3})(\frac{7}{32})}{\frac{1}{4}} = -\frac{1}{6}$$

The result is otherwise obvious also. For, the correlation between the sexes of husband and wife is -1, but that for every other of the remaining five pairs is zero, since the sex of husband or wife will not determine the sex of their children, nor will the sex of one child determine the sex of another. The average value of the correlation between sexes of different members in a household of 4 consisting of a husband, a wife and two children is, therefore, $-\frac{1}{6}$.

The relative change in variance in adopting the household as a unit of sampling in place of an individual is, therefore, $(M-1)\,\rho$ or -50%. In other words, the household will be approximately twice as efficient as a single person for estimating the sex ratio. It is, of course, recognized that households will not all be of the same size and composition, but results from actual observations show that a household will be considerably more efficient as a unit of sampling than a single individual for the character under consideration. The same situation may hold good for characters like the proportion of persons in a family above a certain age. In general, however, ρ will be positive and we may expect the variance to increase with the size of the cluster.

6*a*.4 Estimation from the Sample of the Efficiency of Cluster Sampling

Data for the complete population are seldom available in practice. What is available is only a sample of clusters and the analysis of variance of the elements in the sample. The problem arising in practice is, therefore, to assess the relative efficiency of cluster sampling from the sample data alone.

Let the sample consist of n clusters. Then the analysis of variance for the sample will take the form of Table 6.5.

TABLE 6.5

Analysis of Variance for Sample

Source of Variation		Degrees of Freedom	Mean Square
Between clusters	..	$n-1$	$\frac{1}{n-1}\sum^{n} M\,(\bar{y}_{i.}-\bar{y}_{n.})^2 = Ms_b^2$
Within clusters	..	$n(M-1)$	$\frac{1}{n(M-1)}\sum_{i}^{n}\sum_{j=1}^{M}(y_{ij}-\bar{y}_{i.})^2 = \bar{s}_w^2$
Total sample	..	$nM-1$	$\frac{1}{nM-1}\sum_{i}^{n}\sum_{j=1}^{M}(y_{ij}-\bar{y}_{n.})^2 = s^2$

In a random sample of clusters, s_b^2 and $\bar{s}_w^2$ will provide unbiased estimates of the corresponding values in the population, *viz.*, S_b^2 and $\bar{S}_w^2$. s^2 will not, however, be an unbiased estimate of S^2, since the elements on which it is based cannot be considered to be a simple random sample of elements from the population of NM units. An unbiased estimate of S^2 is, however, easily obtained from (19) by substituting for S_b^2 and $\bar{S}_w^2$ the values s_b^2 and $\bar{s}_w^2$. We thus have

$$\text{Est. } S^2 = \frac{(N-1)\, M s_b^2 + N\,(M-1)\,\bar{s}_w^2}{NM-1} \tag{41}$$

Hence substituting from (41) for Est. (S^2) and writing s_b^2 for S_b^2 in (10), we obtain

$$\begin{aligned} &\text{Est. (Relative Efficiency)} \\ &= \frac{(N-1)\, M s_b^2 + N\,(M-1)\,\bar{s}_w^2}{(NM-1)\, M s_b^2} \\ &\cong \frac{1}{M} + \left(\frac{M-1}{M}\right)\frac{\bar{s}_w^2}{M s_b^2} \end{aligned} \tag{42}$$

for large N.

Example 6.2

Table 6.6 gives the analysis of variance of the area under wheat for a sample of 44 clusters selected from 11 different villages in the Meerut District (India). Four clusters were selected from each of the 11 villages and each cluster consisted of 8 consecutive survey numbers. Estimate the relative efficiency of the cluster as a unit of sampling compared with the individual survey number. The total number of clusters in a village may be assumed to be large.

On substituting the values from Table 6.6 in (41), we obtain for large N,

$$\begin{aligned} \text{Est. } S^2 &\cong s_b^2 + \frac{M-1}{M}\,\bar{s}_w^2 \\ &= \frac{251\cdot 4}{8} + \frac{7}{8}\,(112.8) \\ &= 130\cdot 1 \end{aligned}$$

Hence

Est. (Relative Efficiency)

$$= \frac{130 \cdot 1}{251 \cdot 4}$$

$$= 0 \cdot 52$$

TABLE 6.6

Analysis of Variance of Area under Wheat

Source of Variation	Degrees of Freedom	Mean Square
Between villages	10	$290 \cdot 1$
Between clusters within villages	33	$251 \cdot 4 = Ms_b^2$
Between survey numbers within clusters ..	308	$112 \cdot 8 = \bar{s}_w^2$
Total ..	351	

6a.5 Relationship between the Variance of the Mean of a Single Cluster and its Size

So far we have considered the relative efficiency of clusters and elements as sampling units. The more general problem is that of estimating the variance of the estimated character from a sample of clusters of any size, given the variance of an equivalent sample of clusters of a particular size. This is possible only if we know the relationship between the mean square between means of clusters of given size S_b^2 and M. Several attempts have been made to work out such a relationship. The first one was due to Fairfield Smith (1938).

He argued that:

> If the cluster were to consist of a random sample of elements, S_b^2 would be equal to S^2/M.

Owing to the fact, however, that for most populations encountered in practice elements of a cluster will be positively correlated, clusters will differ in their average values more than when they are composed of randomly selected elements. S_b^2 will, therefore, ordinarily exceed S^2/M.

He proposed the following relationship:

$$S_b^2 = \frac{S^2}{M^g} \tag{43}$$

where g is a constant, less than 1, to be calculated from the sample. He found the relationship to be satisfactory on yield data from uniformity trials for different size plots.

Example 6.3

Table 6.7 shows the values of the mean squares between plots within fields for plots of five different sizes, *viz.*, equilateral triangles of sides (*a*) 33′, (*b*) 25′, (*c*) 15′, (*d*) 10′ and (*e*) 5′ each. The data relate to the crop-cutting survey on wheat conducted in Kangra District (India) during the year 1945–46. Altogether 76 fields were selected for the survey and in each, 10 plots, two of each of the above sizes, were marked at random.

TABLE 6.7

Yield Survey on Wheat, 1945–46 (Kangra)

Values of Mean Squares between Plots within Fields for Plots of Different Sizes

Size of Plot M (Sq.ft.)	Mean Square between Plots within Fields (Md./Acre)^2	
	Observed	Fitted
471·5	0·51	0·56
270·6	0·83	0·75
97·4	1·21	1·26
43·3	2·14	1·91
10·8	3·63	3·88

Examine whether the linear relationship

$$\log S_b^2 = \log S^2 - g \log M$$

proposed by Fairfield Smith describes the data adequately.

The equation obtained by fitting the linear curve by the method of least squares to the data observed is given by

$$\log S_b^2 = 1\cdot117 - 0\cdot511 \log M$$

The theoretical values of S_b^2 as calculated from this equation are given in col. 3 of Table 6.7. It will be seen that the fit is satisfactory, showing that Fairfield Smith's law adequately describes the data.

Fairfield Smith's law, however, leads to one logical difficulty. On the assumption that the total mean square between elements in the population is known and the mean square between means (per element) in clusters of size M is given by the relationship proposed above, an expression for the within cluster mean square can be derived directly from (19). We get

$$\bar{S}_w^2 = \frac{(NM-1)\,S^2 - (N-1)\,M\,\dfrac{S^2}{M^g}}{N\,(M-1)} \tag{44}$$

Equation (44) shows that variability within clusters is a function of N, the size of the finite population, although strictly it should have been independent of it. For N large, however, $\bar{S}_w^2$ becomes independent of N. being given by

$$\bar{S}_w^2 = \frac{M}{M-1}\,S^2\,(1 - M^{-g}) \tag{45}$$

where S^2 will now represent the total mean square in the infinite population of which the finite population is itself a sample. Equation (45) also shows that if we regard the population itself as a single cluster and M is consequently very large, the within cluster variance $\bar{S}_w^2$ will approach S^2 as expected.

If instead of assuming the relationship given by (43) we assume the one given by (45) which satisfies the condition of $\bar{S}_w^2$ being independent of the size of the population, the expression for the mean square between clusters can be written as follows:

$$S_b^2 = \frac{1}{M\,(N-1)}\left\{(NM-1)\,S^2 - N\,(M-1)\,\frac{M}{M-1}\,(1 - M^{-g})\,S^2\right\}$$

$$= \frac{S^2}{M\,(N-1)}\left\{\frac{NM}{M^g} - 1\right\} \tag{46}$$

which in the limit tends to S^2/M^g. S_b^2 now depends on N and in fact increases with it when $g < 1$, which is logical, since the variance of cluster means for clusters of a given size when clusters are widely separated should be larger than the variance observed when they are close together.

Jessen (1942) showed that although Fairfield Smith's original formula, namely (43), describes the yield data extremely well and the refinement suggested by him, namely (46), even improves the relationship, most economic characters relating to farm data follow a slightly different law. He postulated that the mean square among elements within a cluster is a monotone increasing function of the size of the cluster given by

$$\bar{S}_w^2 = aM^b \qquad (b > 0) \tag{47}$$

where a and b are constants to be evaluated from the data. The same relationship was independently suggested by Mahalanobis (1940). Consequently, assuming this law to hold for the mean square within clusters, the expression for the mean square between cluster means is obtained as shown below:

$$S_b^2 = \frac{(NM-1)\,S^2 - N(M-1)\,aM^b}{M(N-1)} \tag{48}$$

The constants S^2, a and b are evaluated from the data. For this purpose, we require: (1) an estimate of the mean square among elements in the population, and (2) an estimate of the mean squares between elements within clusters for at least two values of M. If we regard the total population as a single cluster containing NM elements so that

$$S^2 = a\,(NM)^b$$

then we have

$$S_b^2 = \frac{(NM-1)\,a\,(NM)^b - N(M-1)\,aM^b}{M(N-1)} \tag{49}$$

The above relationship now depends on only two constants and can, therefore, be estimated from the variance among elements in the population and the variance within clusters for any one value of M. Hendricks (1944), however, pointed out that the law may not hold good for large sizes of clusters. This was also the finding of Asthana (1950), who has fitted Jessen's law to describe the mean square within clusters for acreage under wheat for a large number of villages. He found that the observed value of $\bar{S}_w^2$ when the sampling unit is formed by the whole of the population (village) was consistently, though only slightly, below the fitted line showing that the law probably did not hold good for large sizes of clusters. When the law was fitted to the within cluster mean squares corresponding to sizes 2, 4, 8 and 16 only, that is excluding the cluster formed by the whole population, he found that the fit improved.

Example 6.4

Fit Jessen's law to the within cluster mean square values for clusters of sizes 2, 4, 8 and 16 survey numbers and the one formed by all the survey numbers in the village, given in Table 6.8.

TABLE 6.8

Values of $\bar{S}_w^2$ in Clusters of Different Sizes (Acres)^2

M	Observed Values	Fitted Values
2	78·10	81·53
4	84·28	84·25
8	88·92	87·05
16	93·50	89·95
$NM = 1176$	108·33	110·22

The clusters were formed by grouping consecutive survey numbers in a village. The character under study was the area under wheat. Jessen's law as fitted to these values will be found to be given by

$$\log \bar{S}_w^2 = 1{\cdot}897 + 0{\cdot}0473 \log M$$

The theoretical values as calculated from this law are given beside the observed values in Table 6.8. It will be seen that the law adequately describes the data.

6*a*.6 Optimum Unit of Sampling and Multipurpose Surveys

We have seen that ordinarily cluster sampling will lead to loss of precision. On the other hand, cluster sampling helps to reduce the cost of a survey. In this section we shall consider the problem of determining the optimum size of cluster which will provide the maximum information for the funds available.

We shall assume that clusters are equal in size. The cost of a survey based on a sample of n clusters will, apart from over-head costs on planning and analysis, be made up of: (*a*) costs due to the time spent on enumerating all the elements in the sample, nM in number, including the time spent on travelling from one element to another within clusters and costs on transport within clusters, and (*b*) costs due to the time spent on travelling between clusters and the cost of transportation between clusters. The cost of a survey can, therefore, be expressed as a sum of two components one of which is proportional to the number of elements in the sample and the other proportional to the distance to be travelled between clusters, *i.e.*,

$$C = c_1 nM + c_2 d$$

where c_1 represents the cost of enumerating an element including the travel cost from one element to another within the cluster, c_2 that of travelling a unit distance between clusters and d the distance between clusters. It has been shown empirically that the minimum distance between n points located at random is proportional to $n^{\frac{1}{2}} - n^{-\frac{1}{2}}$ (Mahalanobis, 1940). Jessen by means of experimental work has shown that the approximation $n^{\frac{1}{2}}$ works well in practice (1942). The equation for the cost of a survey can, therefore, be expressed as

$$C = c_1 nM + c_2 n^{\frac{1}{2}} \tag{50}$$

where c_2 will now be proportional to the cost of travelling a unit distance.

We have already seen that if the variance within clusters is assumed to follow Jessen's law, then the variance of the estimated mean per element based on a sample of n clusters of size M each is given by

$$V(\bar{y}_{n.}) = \frac{N-n}{N} \cdot \frac{S_b^2}{n} \tag{51}$$

where S_b^2 is as defined in (48). Substituting in (51) for S_b^2 the value from (48), we have

$$V(\bar{y}_{n.}) \cong \frac{1}{n}\left\{\mathrm{S}^2 - (M-1)\,aM^{b-1}\right\} \tag{52}$$

where the finite multiplier is ignored. The problem is to choose n and M so that the variance given by (52) is minimized for a specified cost. We will give the solution as it was first presented by Jessen and then attempt an algebraic solution.

The investigation carried out by Jessen related to a survey of farm facts concerning number of livestock of different types, acreage and yield of corn and oats and income and expenditure on the farm. Samples of seven different sampling units were taken. The sampling units considered were (1) the individual farm, (2) quarter section, S_4, (3) half section, S_2, (4) full section, S, (5) two adjacent sections, $2S$, (6) 4 adjacent sections, $4S$ and (7) 36 sections, $36S$. Using the equation of the cost function, Jessen calculated for each of the different sampling units the total number of clusters n which could be covered for different combinations of two different levels of the total expenditure, three different values of c_1 and two different values of c_2. The two levels of the total expenditure specified were \$ 1000 and \$ 2000 and the values of c_1 assumed were proportional to 1, 4 and 8 and those of c_2 to 1 and 2·5. Table 6.9, reproduced from Jessen's paper (1942), shows the numbers of sampling units which could be covered under each of the several cost situations. Substituting the value of n thus obtained in the equation for the variance, namely (52), he calculated for each of the 7 different sampling units the percentage standard errors of each of the 18 items under study. The results relating to the cost situation in which the total expenditure was fixed at \$ 1000 and c_1 is proportional to 1 and c_2 proportional to 1 are reproduced in Table 6.10. It will be seen that for all except two characters, *viz.*, number of sheep and number of eggs, the quarter section has yielded about the lowest standard error showing that, for the specified budget and c_1 and c_2 as given, S_4 is the optimum size of the sampling unit for the collection of farm facts.

Jessen made similar calculations for different cost situations. Results obtained on six sampling units S_4, S_2, S, $2S$, $4S$ and $36S$ are summarised in Table 6.11. Thus, row 1 of the table shows that when C is equal to \$ 1000 and $c_1 = 1$ and $c_2 = 1$, for 10 out of 18 characters under study, the half section S_2 will be the optimum unit of sampling. Row 3 of the table similarly shows that for the same total budget and c_2 remaining the same, but c_1 increased to 8 times its value, the quarter section S_4 would be the optimum unit for 16 out of the 18 characters. The result indicated that the size of the optimum unit of sampling decreases as c_1, the cost of enumeration, increases. This is confirmed by examination of other parts of the table. Similarly, the table shows that the increase in c_2 calls for the use of larger sampling units. We also see that a large budget requires small clusters.

From the study of similar results for 1939, Jessen recommended the use of the quarter section as the optimum unit of sampling for this kind of survey involving the collection of information on several items. This is an important finding since it points to the possibility of obtaining information on several related items in the compass of a single survey using the same optimum unit of sampling. This is also the experience in India with regard to the use of the village as the unit of sampling in agricultural surveys which has the additional advantage of being administratively convenient (Sukhatme, 1950; Sukhatme and Panse, 1951). The degree of accuracy attained necessarily varies from character to character in a given sample of clusters of the optimum size, but this can be adjusted using different intensities of sampling for different groups of items in the questionnaire. Thus, we may divide the questionnaire into three parts, information on part 1 to be collected from only half the total sample, that on part 2 from say three-fourths of the sample and on part 3 from the entire sample. The items may be so grouped that information on all will be available with about the desired precision. Such sample surveys which include within their scope the collection of information on more than one item are called *multipurpose* surveys.

We shall now give an algebraic solution of the problem originally due to Cochran (1948) of choosing M and n so that the variance given by (52) is minimized for a given value of the total cost, say $C = C_0$.

We form the function

$$\phi = V(\bar{y}_{n..}) + \mu\,(c_1 nM + c_2 n^{\frac{1}{2}} - C_0) \tag{53}$$

where μ is the Lagrangian undetermined constant. We next differentiate with respect to n, M and μ and equate the results to zero. Thus, on differentiating with respect to n and noting that

$$\frac{\partial V}{\partial n} = -\frac{V}{n} \tag{54}$$

we get

$$-\frac{V}{n} + \mu\,(c_1 M + \tfrac{1}{2} c_2 n^{-\frac{1}{2}}) = 0 \tag{55}$$

Similarly, on differentiating with respect to M and equating the result to zero, we have

$$\frac{\partial V}{\partial M} + \mu c_1 n = 0 \tag{56}$$

And finally differentiating with respect to μ and equating to zero, we have

$$c_1 nM + c_2 n^{\frac{1}{2}} = C_0 \tag{57}$$

On eliminating μ from equations (55) and (56), we obtain

$$-\frac{\partial V}{\partial M} \cdot \frac{M}{V} = \frac{1}{1 + \left(\dfrac{c_2}{2c_1 M n^{\frac{1}{2}}}\right)} \tag{58}$$

Now solving equation (57) as a quadratic in $n^{\frac{1}{2}}$, we get

$$n^{\frac{1}{2}} = \frac{-c_2 + (c_2^2 + 4c_1 C_0 M)^{\frac{1}{2}}}{2c_1 M} \tag{59}$$

On substituting for $n^{\frac{1}{2}}$ in (58) and simplifying the algebra, we obtain

$$\frac{M}{V} \cdot \frac{\partial V}{\partial M} = -1 + \left(1 + \frac{4c_1C_0M}{c_2{}^2}\right)^{-\frac{1}{2}} \tag{60}$$

Now it can be seen from (52) that

$$\frac{1}{V} \cdot \frac{\partial V}{\partial M}$$

is independent of n. Equation (60) can, therefore, be solved directly for M. An explicit expression for M is, however, difficult to obtain and the solution has, therefore, to be obtained by trial and error method. On substituting the value of M so obtained back in (59), we obtain the optimum value of n.

Since V decreases as M increases, we may expect the left-hand side of (60) to be approximately constant whatever the value of M. An examination of (60) also shows that the left-hand side is independent of the cost factors while the right-hand side involves M only in combination with the cost factors. It follows therefore that M will respond to the variation in c_1, c_2 and C in such a way that $c_1CM/c_2{}^2$ is approximately constant. It follows that M will be smaller if (1) c_1 increases, *i.e.*, the cost of enumerating an element increases; (2) c_2 decreases, *i.e.*, travel becomes cheaper; and (3) C is large, *i.e.*, the amount of money available for the survey is large. The algebraic solution thus confirms the calculations deduced from the actual data reproduced in Tables 6.9–6.11.

TABLE 6.9

Numbers of Sampling Units which can be Covered, given Several Cost Situations, Two Expenditure Levels, and Seven Different Sampling Units

Unstratified Sample in the State of Iowa

Sampling Unit	No. of Farms/ Sampling Unit*	Mileage at 2¢/Mile			Mileage at 5¢/Mile		
		Length of Farm Visit			Length of Farm Visit		
		15 Min.	60 Min.	120 Min.	15 Min.	60 Min.	120 Min.
		A. Total Expenditure of $1000					
Individual farm	1·000	1644	650	371	1088	517	315
Quarter section	0·914	1745	699	401	1140	551	339
Half section	1·828	1073	392	218	764	336	192
Section	3·656	624	213	116	475	186	105
Two sections	7·312	347	113	60	278	102	56
Four sections	14·624	187	59	31	156	54	29
Thirty-six sections	131·616	21	7	4	17	6	3
		B. Total Expenditure of $2000					
Individual farm	1·000	4012	1452	803	2886	1223	712
Quarter section	0·914	4293	1569	871	3057	1314	769
Half section	1·828	2494	852	462	1900	744	421
Section	3·656	1388	451	241	1112	407	225
Two sections	7·312	749	235	124	623	217	118
Four sections	14·624	396	121	63	338	113	61
Thirty-six sections	131·616	44	14	7	38	13	7

* Computed from the sample survey data.

TABLE 6.10

Relative Standard Errors (% of Item Means per Farm) Estimated for Samples of Different Sampling Units and Taken at Random within the State, 1938

(Expenditure of $1000, 15-minute Questionnaire and 2c per Mile)

Items	Sampling Unit						
	Individual Farm	S_4	S_2	S	$2S$	$4S$	$36S$
1. Number of swine	2·67	2·82	2·74	2·90	3·36	4·11	9·99
2. Number of horses	1·83	1·93	1·87	1·98	2·27	2·80	6·87
3. Number of sheep	9·61	9·76	8·80	3·16	7·74	7·44	7·44
4. Number of chickens ..	1·61	1·70	1·66	1·78	2·07	2·57	6·34
5. Number of eggs yesterday ..	3·17	3·21	2·90	2·69	2·55	2·45	2·45
6. Number of cattle	2·55	2·67	2·55	2·65	2·98	3·62	8·66
7. Number of cows milked ..	1·98	2·07	2·00	2·09	2·37	2·88	6·79
8. Number of gallons of milk ..	2·34	2·45	2·32	2·39	2·64	3·15	7·17
9. Dairy product receipts ..	2·99	3·11	2·93	2·97	3·24	3·79	8·55
10. Number of farm acres ..	1·54	1·63	1·57	1·64	1·87	2·28	5·58
11. Number of corn acres ..	1·95	2·06	1·98	2·08	2·37	2·87	6·88
12. Number of oat acres ..	2·36	2·59	2·66	3·05	3·78	4·91	12·76
13. Corn yield	·82	·90	·94	1·09	1·36	1·78	4·73
14. Oat yield	·84	·88	·84	·86	·96	1·15	2·71
15 Commercial feed expenditures	6·23	7·06	7·60	9·14	11·78	15·71	43·07
16. Total expenditures, operator	3·96	4·36	4·51	5·21	6·48	8·46	22·36
17. Total receipts, operator ..	3·16	3·49	3·64	4·23	5·29	6·93	18·39
18. Net cash income, operator ..	3·54	3·82	3·84	4·26	5·13	6·57	16·82

TABLE 6.11

Summary of Sampling Unit Efficiencies

Number of Items Most Efficiently Estimated by the Six-Grid Sampling Units, 1938 and 1939

Expenditure, Mileage Rate and Questionnaire Length		Sampling Unit					
		S_4	S_2	S	$2S$	$4S$	$36S$
		Expenditure of $1000					
I. 2e/15 min.	1938 ..	6	10	..	..	1	1
	1939 ..	6½	8½	1	..	2	2
II. 2e/60 min.	1938 ..	13	3	..	..	1	1
	1939 ..	14	2	..	..	2	2
III. 2e/120 min.	1938 ..	16	..	..	..	1	1
	1939 ..	16	..	..	..	2	2
IV. 5e/15 min.	1938 ..	1	12½	2½	..	1	1
	1939 ..	4	9	3	..	2	2
V. 5e/60 min.	1938 ..	6	10	..	..	1	1
	1939 ..	7½	8½	..	..	2	2
VI. 5e/120 min.	1938 ..	11½	4½	..	..	1	1
	1939 ..	12	4	..	..	2	2
		Expenditure of $2000					
VII. 2e/15 min.	1938 ..	7	9	..	..	1	1
	1939 ..	8	8	..	..	2	2
VIII. 2e/60 min.	1938 ..	16	..	..	..	1	1
	1939 ..	15	1	..	..	2	2
IX. 2e/120 min.	1938 ..	16	..	..	..	1	1
	1939 ..	16	1	..	..	2	2
X. 5e/15 min.	1938 ..	5	11	..	..	1	1
	1939 ..	6	8	2	..	2	2
XI. 5e/60 min.	1938 ..	12½	3½	..	..	1	1
	1939 ..	12	4	..	..	2	2
XII. 5e/120 min.	1938 ..	12½	3½	..	..	1	1
	1939 ..	14	2	..	..	2	2

B. UNEQUAL CLUSTERS

6*b*.1 Estimates of the Mean and their Variances

Let the *i*-th cluster consist of M_i elements ($i=1, 2, \ldots, N$) and let $M_0 = \sum_{i=1}^{N} M_i$ denote the total number of elements and $\bar{M}=M_0/N$ the average number of elements per cluster in the population. Then the mean of the character per element in the *i*-th cluster will be given by

$$\bar{y}_{i.} = \frac{1}{M_i} \sum_{j=1}^{M_i} y_{ij} \tag{61}$$

and the mean per element in the population by

$$\bar{y}_{..} = \frac{\sum_{i=1}^{N} \sum_{j=1}^{M_i} y_{ij}}{\sum_{i=1}^{N} M_i} = \frac{\sum_{i=1}^{N} M_i \bar{y}_{i.}}{M_0} . \tag{62}$$

Several estimates of the population value of the mean per element can be formed from a random sample of *n* clusters. We shall first consider the simplest, *viz.*, the simple arithmetic mean of the cluster means given by

$$\bar{y}_{n.} = \frac{1}{n} \sum^{n} \bar{y}_{i.} \tag{63}$$

It is easy to see that this estimate will not give an unbiased estimate of the population value, for

$$\begin{aligned} E(\bar{y}_{n.}) &= \frac{1}{n} \sum^{n} E(\bar{y}_{i.}) \\ &= \frac{1}{n} \sum^{n} \left\{ \frac{1}{N} \sum_{h=1}^{N} \bar{y}_{h.} \right\} \\ & \quad \frac{1}{N} \sum_{h=1}^{N} \bar{y}_{h.} \\ &= \bar{y}_{N} . \qquad\qquad (64) \\ &\neq \bar{y}_{..} \end{aligned}$$

unless M_i and $\bar{y}_{i.}$ are uncorrelated. It is likely, however, that for large n and for a population for which M_i's do not appreciably vary from one cluster to another, this estimate may not be materially biased.

Since $\bar{y}_{n.}$ is a biased estimate, the error in $\bar{y}_{n.}$ will consist of two components: one arising from the sampling variations about its own mean, *viz.*, the unweighted mean of the cluster means in the population; and the other arising from the bias component. The expected value of the square of the total error in $\bar{y}_{n.}$ is called the mean square error. To evaluate it, we write

$$\bar{y}_{n.} - \bar{y}_{..} = \bar{y}_{n.} - \bar{y}_{N.} + \bar{y}_{N.} - \bar{y}_{..} \tag{65}$$

where $\bar{y}_{N.}$ is the simple arithmetic mean of the cluster means, given by

$$\bar{y}_{N.} = \frac{1}{N}\sum_{i=1}^{N} \bar{y}_{i.}$$

Squaring both sides of (65) and taking expectations, we obtain

$$\begin{aligned} M.S.E.\,(\bar{y}_{n.}) &= V(\bar{y}_{n.}) + (\text{bias})^2 \\ &= \frac{N-n}{N} \cdot \frac{S_b^2}{n} + (\bar{y}_{N.} - \bar{y}_{..})^2 \end{aligned} \tag{66}$$

where S_b^2 is defined in (6), namely,

$$S_b^2 = \frac{1}{N-1}\sum_{i=1}^{N} (\bar{y}_{i.} - \bar{y}_{N.})^2$$

An unbiased estimate can also be formed. The simplest would be the arithmetic mean based on cluster totals given by

$$\bar{y}_{n.}' = \frac{1}{n\bar{M}}\sum^{n} M_i \bar{y}_{i.} \tag{67}$$

It is easy to see that it is unbiased, for

$$E(\bar{y}_{n.}') = \frac{1}{n\bar{M}}\sum^{n} E(M_i \bar{y}_{i.})$$

$$= \frac{1}{n\bar{M}} \cdot \frac{n}{N} \sum_{i=1}^{N} M_i \bar{y}_{i.}$$

$$= \bar{y}_{..} \tag{68}$$

The sampling variance of this estimate can be written as

$$V(\bar{y}_{n.}') = \frac{N-n}{N} \cdot \frac{1}{n} \mathbf{S}_b'^2 \tag{69}$$

where

$$\mathbf{S}_b'^2 = \frac{1}{N-1} \sum_{i=1}^{N} \left(\frac{M_i \bar{y}_{i.}}{\bar{M}} - \frac{1}{N} \sum_{i=1}^{N} \frac{M_i \bar{y}_{i.}}{\bar{M}} \right)^2$$

$$= \frac{1}{N-1} \sum_{i=1}^{N} \left(\frac{M_i \bar{y}_{i.}}{\bar{M}} - \bar{y}_{..} \right)^2 \tag{70}$$

It will be noticed that the variance of $\bar{y}_{n.}'$ depends upon the variation of the product $M_i\bar{y}_{i.}$, and is, therefore, likely to be larger than that of $\bar{y}_{n.}$, unless M_i and $\bar{y}_{i.}$ vary in such a way that their product is almost constant.

A third estimate which is biased but consistent is given by

$$\bar{y}_{n.}'' = \frac{\overset{n}{\Sigma}\, M_i \bar{y}_{i.}}{\overset{n}{\Sigma}\, M_i} \tag{71}$$

It is a weighted mean of the cluster means and is the ratio of two random variables. We have already seen that this estimate is a biased estimate, but is consistent, the bias decreasing with the increase in n. A first approximation to the variance of this estimate is given by replacing y_i by $M_i\bar{y}_{i.}$ and x_i by M_i in equation (30) of Chapter IV. We obtain

$$V_1(\bar{y}_{n.}'') = \frac{N-n}{Nn} \cdot \frac{1}{N-1} \sum_{i=1}^{N} \frac{M_i^2}{\bar{M}^2} (\bar{y}_{i.} - \bar{y}_{..})^2$$

$$= \frac{N-n}{Nn} \mathbf{S}_b''^2 \tag{72}$$

where

$$S_b''^2 = \frac{1}{N-1} \sum_{i=1}^{N} \frac{M_i^2}{\bar{M}^2} (\bar{y}_{i.} - \bar{y}_{..})^2 \tag{73}$$

which, as we have seen, is a satisfactory approximation to the actual variance provided n is large.

The variance of the ratio estimate is smaller than that based on the simple arithmetic mean of the cluster totals provided ρ is larger than $C.V. (M_i)/2\ C.V. (M_i\bar{y}_{i.})$ and consequently the estimate $\bar{y}_{n.}''$ is expected to be more efficient than $\bar{y}_{n.}'$ in large samples, whenever M_i and $M_i\bar{y}_{i.}$ are highly correlated.

6*b*.2 Probability Proportional to Cluster Size: Estimate of the Mean and its Variance

The basic theory of sampling with varying probabilities of selection has been given in Sections 2*b*.2, 2*b*.3, 2*b*.4 and 2*b*.5 of Chapter II. In this section we shall give its application to cluster sampling.

Let P_i denote the probability of selecting the i-th cluster ($i = 1, 2, \ldots, N$) at the first draw, $\sum\limits_{i=1}^{N} P_i$ being 1, and define a variate z by

$$z_{ij} = \frac{M_i y_{ij}}{M_0 P_i} \tag{74}$$

whence

$$\bar{z}_{i.} = \frac{M_i \bar{y}_{i.}}{M_0 P_i}$$

Then it is easily shown that the expected value of $\bar{z}_{i.}$ is equal to the population mean $\bar{y}_{..}$. Denoting the expected value of $\bar{z}_{i.}$ by $\bar{z}_{..}$, we have

$$\begin{aligned} \bar{z}_{..} &= \sum_{i=1}^{N} P_i \bar{z}_{i.} \\ &= \sum_{i=1}^{N} P_i \frac{M_i}{M_0} \frac{\bar{y}_{i.}}{P_i} \\ &= \bar{y}_{..} \end{aligned} \tag{75}$$

It follows that

$$E(\bar{z}_{n.}) = \bar{z}_{..} = \bar{y}_{..} \tag{76}$$

When the selection probability is proportional to the size of the cluster, in other words, when $P_i = M_i/M_0$, the variate z becomes identical with y, and $\bar{z}_{n.} = \bar{y}_{n.}$. We thus reach an important result, that a simple arithmetic mean of the cluster means, under a system of sampling with probability proportional to size of cluster, gives an unbiased estimate of the population mean $\bar{y}_{..}$.

To obtain the sampling variance of $\bar{z}_{n.}$, we proceed exactly step by step as in Section 2*b*.2 and reach the same expression as in (136), namely,

$$V(\bar{z}_{n.}) = \frac{\sigma_{bz}^2}{n} \tag{77}$$

where σ_{bz}^2 stands for the variance of a cluster mean, defined by

$$\sigma_{bz}^2 = \sum_{i=1}^{N} P_i (\bar{z}_{i.} - \bar{z}_{..})^2 \tag{78}$$

The estimate of the sampling variance of $\bar{z}_{n.}$ is also given by the same expression as shown in (143) in Section 2*b*.3, namely,

$$\text{Est. } V(\bar{z}_{n.}) = \frac{s_{bz}^2}{n} \tag{79}$$

where s_{bz}^2 is the mean square between cluster $\bar{z}_{i.}$'s in the sample, defined by

$$s_{bz}^2 = \frac{1}{n-1} \sum^{n} (\bar{z}_{i.} - \bar{z}_{n.})^2 \tag{80}$$

When $P_i = M_i/M_0$, we have $\bar{z}_{n.} = \bar{y}_{n.}$,

$$V(\bar{z}_{n.}) = \frac{\sigma_b^2}{n} \tag{81}$$

where

$$\sigma_b^2 = \sum_{i=1}^{N} \frac{M_i}{M_0} (\bar{y}_{i.} - \bar{y}_{..})^2 \tag{82}$$

and

$$\text{Est. } V(\bar{z}_{n.}) = \frac{s_b^2}{n} \tag{83}$$

where

$$s_b^2 = \frac{1}{n-1} \sum^{n} (\bar{y}_{i.} - \bar{y}_{n.})^2 \tag{84}$$

6*b*.3 Probability Proportional to Cluster Size: Efficiency of Cluster Sampling

It can be shown that the relative change in variance with a cluster as a unit of sampling in place of an element under the system of sampling with probability proportional to size of cluster is given by an expression similar to that in the case of equal clusters, *viz*., $(\bar{M} - 1)\,\rho$. To evaluate ρ, we start from equations (13) and (14) and write

$$\rho = \frac{E\,\{(y_{ij} - \bar{y}_{i.})\,(y_{ik} - \bar{y}_{i.})\} + E\,(\bar{y}_{i.} - \bar{y}_{..})^2}{E\,(y_{ij} - \bar{y}_{..})^2} \tag{85}$$

By definition

$$E\,(\bar{y}_{i.} - \bar{y}_{..})^2 = \sigma_b^2 \tag{86}$$

Further, let

$$E\,\{(y_{ij} - \bar{y}_{i.})^2 \mid i\} = \sigma_i^2 \tag{87}$$

and

$$E\,(y_{ij} - \bar{y}_{..})^2 = \sigma^2 \tag{88}$$

To obtain the expected value of the first term in the numerator of (85), we use the identity

$$\left\{\sum_{j=1}^{M_i} (y_{ij} - \bar{y}_{i.}) \mid i\right\}^2 \equiv 0 \equiv \sum_{j=1}^{M_i} (y_{ij} - \bar{y}_{i.})^2$$

$$+ \sum_{j \neq k = 1}^{M_i} (y_{ij} - \bar{y}_{i.})\,(y_{ik} - \bar{y}_{i.})$$

which can be written as

$$\sum_{j=1}^{M_i} \{(y_{ij} - \bar{y}_{i.})^2 \mid i\} + M_i\,(M_i - 1)\;E\,\{(y_{ij} - \bar{y}_{i.})\,(y_{ik} - \bar{y}_{i.}) \mid i\} \equiv 0$$

Hence

$$E\{(y_{ij}-\bar{y}_{i.})(y_{ik}-\bar{y}_{i.})\mid i\} = -\frac{\sigma_i^2}{M_i-1} \tag{89}$$

Taking the expectation of the expressions on both sides of (89) for variation in *i*, we obtain

$$E\{(y_{ij}-\bar{y}_{i.})(y_{ik}-\bar{y}_{i.})\} = -\sum_{i=1}^{N} P_i \cdot \frac{\sigma_i^2}{M_i-1} \tag{90}$$

Assuming σ_i^2 to be constant and equal to σ_w^2, we have

$$E\{(y_{ij}-\bar{y}_{i.})(y_{ik}-\bar{y}_{i.})\} = -\frac{\sigma_w^2}{N\bar{M}}\sum_{i=1}^{N}\frac{M_i}{M_i-1} \tag{91}$$

Now the expression

$$\frac{1}{N}\sum_{i=1}^{N}\frac{M_i}{M_i-1}$$

may be satisfactorily approximated by $\bar{M}/(\bar{M}-1)$. We therefore have

$$E\{(y_{ij}-\bar{y}_{i.})(y_{ik}-\bar{y}_{i.})\} = -\frac{\sigma_w^2}{\bar{M}-1} \tag{92}$$

Hence, substituting from (86), (88) and (92) in (85), we have

$$\rho = \frac{-\dfrac{\sigma_w^2}{\bar{M}-1}+\sigma_b^2}{\sigma^2} \tag{93}$$

Also

$$\sigma^2 = \sigma_w^2 + \sigma_b^2 \tag{94}$$

Hence, eliminating σ_w^2 from equations (93) and (94), we have

$$\bar{M}\sigma_b^2 = \sigma^2\{1+(\bar{M}-1)\rho\} \tag{95}$$

It follows that the sampling variance of the mean of *n* clusters (on an element basis) selected with probability proportional to their size is given by

$$V(\bar{y}_{n.}) = \frac{\sigma_b^2}{n}$$

$$= \frac{1}{n\bar{M}} \sigma^2 \left\{ 1 + (\bar{M} - 1) \rho \right\} \tag{96}$$

If a simple random sample of $n\bar{M}$ elements had been selected independently with replacement, the variance would have been

$$V(\bar{y}_{n\bar{M}}) = \frac{\sigma^2}{n\bar{M}} \tag{97}$$

so that the relative change in variance is given by

$$\frac{(96) - (97)}{(97)} = (\bar{M} - 1) \rho \tag{98}$$

6b.4 Probability Proportional to Cluster Size: Relative Efficiency of Different Estimates

In Section 2b.2 of Chapter II we remarked that the method of sampling with selection probabilities proportional to the size of the clusters may give marked reduction in variance of the estimated means compared to the method of simple random sampling of clusters. In this section we shall make the relevant comparisons.

Of the three estimates appropriate for simple random sampling, namely, $\bar{y}_{n.}$, $\bar{y}_{n.}'$ and $\bar{y}_{n.}''$, it is necessary to consider only the first and the last, since the estimate $\bar{y}_{n.}'$ will generally be less efficient than either $\bar{y}_{n.}$ or $\bar{y}_{n.}''$. We shall, therefore, make comparisons under two heads:

(*A*) that of $\bar{z}_{n.}$ with $\bar{y}_{n.}$, and

(*B*) that of $\bar{z}_{n.}$ with $\bar{y}_{n.}''$.

We shall assume that sampling is carried out with replacement.

(A) Comparison of $\bar{z}_{n.}$ with $\bar{y}_{n.}$

We have seen that $\bar{z}_{n.}$ is an unbiased estimate of $\bar{y}_{..}$ with its variance given by

$$V(\bar{z}_{n.}) = \frac{1}{nN} \sum_{i=1}^{N} \frac{M_i}{\bar{M}} (\bar{y}_{i.} - \bar{y}_{..})^2 \tag{99}$$

The estimate $\bar{y}_{n.}$, on the other hand, is a biased estimate of the population value. In comparing it with $\bar{z}_{n.}$ we have, therefore, to consider its mean square error which, ignoring the finite multiplier and replacing $N - 1$ by N, is given by

$$M.S.E.\,(\bar{y}_{n.}) \cong \frac{1}{nN} \sum_{i=1}^{N} (\bar{y}_{i.} - \bar{y}_{N.})^2 + (\bar{y}_{N.} - \bar{y}_{..})^2 \tag{100}$$

Using the identity

$$\sum_{i=1}^{N} (\bar{y}_{i.} - \bar{y}_{..})^2 \equiv \sum_{i=1}^{N} (\bar{y}_{i.} - \bar{y}_{N.})^2 + N(\bar{y}_{N.} - \bar{y}_{..})^2$$

equation (100) can be rewritten to read

$$M.S.E.\,(\bar{y}_{n.}) \cong \frac{1}{nN} \sum_{i=1}^{N} (\bar{y}_{i.} - \bar{y}_{..})^2 + \left(1 - \frac{1}{n}\right) (\bar{y}_{N.} - \bar{y}_{..})^2 \tag{101}$$

The difference between (101) and (99) works out to

$$M.S.E.\,(\bar{y}_{n.}) - M.S.E.\,(\bar{z}_{n.}) = -\frac{1}{nN} \sum_{i=1}^{N} \left(\frac{M_i}{\bar{M}} - 1\right) (\bar{y}_{i.} - \bar{y}_{..})^2$$

$$+ \left(1 - \frac{1}{n}\right) (\bar{y}_{N.} - \bar{y}_{..})^2 \tag{102}$$

In order to examine this difference we shall group together clusters of the same size and write the first term in (102) as shown below:

$$-\frac{1}{nN}\sum_{i=1}^{N}\left(\frac{M_i}{\bar{M}}-1\right)(\bar{y}_{i.}-\bar{y}_{..})^2$$

$$=-\frac{1}{nN}\sum_{i=1}^{k}\left(\frac{M_i}{\bar{M}}-1\right)\sum_{j=1}^{N_i}(\bar{y}_{j.}-\bar{y}_{..})^2 \qquad (103)$$

where the summation $\sum_j$ is taken over the N_i clusters of size M_i each, and the summation $\sum_i$ is taken over the k groups, where $\sum_{i=1}^{k} N_i = N$. We shall restrict the discussion to the case where $E(\bar{y}_{j.} \mid M_i) = \bar{y}_{..}$. Clearly,

$$\frac{1}{N_i}\sum_{j=1}^{N_i}(\bar{y}_{j.}-\bar{y}_{..})^2$$

will then represent the variance of $\bar{y}_{j.}$ and may be denoted by $V(\bar{y}_{j.} \mid M_i)$. Equation (103) can now be put in the form

$$-\frac{1}{nN}\sum_{i=1}^{N}\left(\frac{M_i}{\bar{M}}-1\right)(\bar{y}_{i.}-\bar{y}_{..})^2$$

$$=-\frac{1}{nN}\sum_{i=1}^{k}N_i\left(\frac{M_i}{\bar{M}}-1\right)V(\bar{y}_{j.} \mid M_i) \qquad (104)$$

and is seen to depend upon the relationship between the variance of the mean of a cluster of given size and the size. If the clusters were randomly formed, the variance will be inversely proportional to the size of the cluster. Owing to the fact, however, that the elements of a cluster will be correlated, the variance will seldom decrease quite as fast. We shall examine the difference for three special cases:

I. $V(\bar{y}_{j.} \mid M_i) =$ a constant, say γ

II. $V(\bar{y}_{j.} \mid M_i) = \dfrac{\gamma}{M_i}$

and

$$\text{III.} \quad V(\bar{y}_{j.} \mid M_i) = \frac{\gamma}{M_i^2}$$

Case I.—

$$V(\bar{y}_{j.} \mid M_i) = \gamma$$

Clearly the value of (104) is zero, giving us

$$M.S.E.(\bar{y}_{n.}) - M.S.E.(\bar{z}_{n.}) = \left(1 - \frac{1}{n}\right)(\bar{y}_{N.} - \bar{y}_{..})^2 \tag{105}$$

which is always positive, so that $\bar{z}_{n.}$ is expected to be more precise than the estimate $\bar{y}_{n.}$ for this case.

Case II.—

$$V(\bar{y}_{j.} \mid M_i) = \frac{\gamma}{M_i}$$

Equation (104) can be approximated by

$$- \frac{1}{nN} \sum_{i=1}^{N} \left(\frac{M_i}{\bar{M}} - 1\right)(\bar{y}_{i.} - \bar{y}_{..})^2$$

$$= - \frac{\gamma}{nN} \sum_{i=1}^{k} N_i \left(\frac{M_i}{\bar{M}} - 1\right) \frac{1}{M_i}$$

$$= - \frac{\gamma}{nN} \sum_{i=1}^{k} N_i \left(\frac{M_i}{\bar{M}} - 1\right) \frac{1}{\bar{M}} \frac{1}{\left(1 + \frac{M_i - \bar{M}}{\bar{M}}\right)}$$

$$= - \frac{\gamma}{nN\bar{M}} \sum_{i=1}^{N} \left(\frac{M_i}{\bar{M}} - 1\right) \left\{1 - \left(\frac{M_i}{\bar{M}} - 1\right) + \left(\frac{M_i}{\bar{M}} - 1\right)^2 - \cdots\right\}$$

$$\cong \frac{\gamma}{nN\bar{M}} \sum_{i=1}^{N} \left(\frac{M_i}{\bar{M}} - 1\right)^2 \tag{106}$$

which is again positive, so that for this case also we can expect $\bar{z}_{n.}$ to be more efficient than $\bar{y}_{n.}$.

Case III.—

$$V(\bar{y}_{i.} \mid M_i) = \frac{\gamma}{M_i^2}$$

It can be shown that, following the approach in Case II, equation (104) can be approximated by

$$-\frac{1}{nN}\sum_{i=1}^{N}\left(\frac{M_i}{\bar{M}}-1\right)(\bar{y}_{i.}-\bar{y}_{..})^2 \cong \frac{2\gamma}{nN\bar{M}^2}\sum_{i=1}^{N}\left(\frac{M_i}{\bar{M}}-1\right)^2 \tag{107}$$

so that for this case also we would expect $\bar{z}_{n.}$ to be more efficient than $\bar{y}_{n.}$.

These results are in fact obvious from an examination of the first term in (102). This term, ignoring the sign, represents the covariance between M_i and $(\bar{y}_{i.} - \bar{y}_{..})^2$. Now, ordinarily $(\bar{y}_{i.}-\bar{y}_{..})^2$ will decrease as M_i increases so that the covariance will be negative and consequently the expression on the right-hand side of (102) will be positive.

(B) Comparison of $\bar{z}_{n.}$ with $\bar{y}_{n.}''$

The mean square error of $\bar{y}_{n.}''$ when the finite multiplier is ignored and n is large is given by

$$M.S.E.\ (\bar{y}_{n.}'') \cong \frac{1}{nN}\sum_{i=1}^{N}\frac{M_i^2}{\bar{M}^2}(\bar{y}_{i.}-\bar{y}_{..})^2 \tag{108}$$

Deducting from this the *M.S.E.* in (99), we obtain

$$M.S.E.\ (\bar{y}_{n.}'') - M.S.E.\ (\bar{z}_{n.})$$

$$= \frac{1}{nN}\sum_{i=1}^{N}\frac{M_i}{\bar{M}}(\bar{y}_{i.}-\bar{y}_{..})^2\left(\frac{M_i}{\bar{M}}-1\right) \tag{109}$$

Restricting again the discussion to the case in which $E(\bar{y}_{i.})=\bar{y}_{..}$, we notice that the relative precision of $\bar{y}_{n.}''$ and $\bar{z}_{n.}$ depends upon the relationship between the variance of $\bar{y}_{i.}$ and M_i.

Case I.—Let

$$V(\bar{y}_{i.} \mid M_i) = \gamma$$

Equation (109) can then be written as

$$M.S.E.\,(\bar{y}_{n.}{''}) - M.S.E.\,(\bar{z}_{n.}) = \frac{\gamma}{nN} \sum_{i=1}^{N} \frac{M_i}{\bar{M}} \left(\frac{M_i}{\bar{M}} - 1\right)$$

$$= \frac{\gamma}{n} V\left(\frac{M_i}{\bar{M}}\right) \qquad (110)$$

It follows that $\bar{z}_{n.}$ is more efficient than $\bar{y}_{n.}{''}$.

Case II.—Let

$$V(\bar{y}_{i.} \mid M_i) = \frac{\gamma}{M_i}$$

The right-hand side of (109) reduces to zero, showing that the two estimates are of equal precision.

Case III.—Let

$$V(\bar{y}_{i.} \mid M_i) = \frac{\gamma}{M_i^2}$$

Equation (109) then takes the form

$$M.S.E.\,(\bar{y}_{n.}{''}) - M.S.E.\,(\bar{z}_{n.})$$

$$= \frac{\gamma}{nN\bar{M}^2} \sum_{i=1}^{N} (M_i - \bar{M}) \left(\frac{1}{M_i}\right)$$

$$= \frac{\gamma}{nN\bar{M}^3} \sum_{i=1}^{N} (M_i - \bar{M}) \left(1 + \frac{M_i - \bar{M}}{\bar{M}}\right)^{-1}$$

$$\cong -\frac{\gamma}{nN\bar{M}^4} \sum_{i=1}^{N} (M_i - \bar{M})^2 \qquad (111)$$

Thus for this case, in contrast to the previous two, the estimate $\bar{z}_{n.}$ is expected to be less efficient than the ratio estimate in simple random sampling.

Example 6.5

Table 6.12 gives the number of villages and the area under wheat in each of 89 administrative areas* in Hapur Subdivision of Meerut District (India), and Table 6.13 gives the analysis of variance on a village basis. It is required to estimate the total area under wheat in the subdivision using an administrative circle as the unit of sampling. We shall assume that a sample of 20 circles is to be selected. Calculate the sampling variance of the estimate of the total area under wheat for each of the following procedures of sampling and estimation:

(*a*) equal probability, mean of the cluster means estimate,

(*b*) equal probability, mean of the cluster totals estimate,

(*c*) equal probability, ratio estimate,

(*d*) probability proportional to the size of the circle, mean of the cluster means estimate.

Also calculate the variance of an equivalent sample with the village as the unit of sampling and compare the relative efficiency of the various methods.

(*a*) *Equal Probability, Mean of the Cluster Means Estimate*

$$M.S.E.\ (M_0\bar{y}_{n..})$$

$$= M_0^2\left\{\frac{N-n}{Nn}\cdot\frac{1}{N-1}\sum_{i=1}^{N}(\bar{y}_{i.}-\bar{y}_{N.})^2+(\bar{y}_{N.}-\bar{y}_{..})^2\right\}$$

$$= (299)^2\left\{\frac{69}{89\times 20}\cdot\frac{1}{88}\,(6499209)+(387\cdot 35-328\cdot 02)^2\right\}$$

$$= 89401\,(2862\cdot 9+3511\cdot 75)$$

$$= 5699\times 10^5 \text{ acres}^2$$

* These areas are known as Patwari circles in the local terminology.

TABLE 6.12

Number of Villages and the Area under Wheat in the Administrative Circles of Hapur

Circle No. (i)	Number of Villages (M_i)	Area under Wheat (Acres) ($M_i\bar{y}_{i.}$)	Circle No. (i)	Number of Villages (M_i)	Area under Wheat (Acres) ($M_i\bar{y}_{i.}$)
1	6	1562	29	2	583
2	5	1003	30	4	1150
3	4	1691	31	3	670
4	5	271	32	2	499
5	4	458	33	4	714
6	2	736	34	4	1081
7	4	1224	35	1	389
8	2	996	36	7	2675
9	5	475	37	3	868
10	1	34	38	2	1412
11	3	1027	39	2	445
12	4	1393	40	5	706
13	3	692	41	2	642
14	1	524	42	4	2050
15	1	602	43	6	2530
16	3	1522	44	1	247
17	4	2087	45	2	421
18	8	2474	46	2	687
19	2	461	47	3	941
20	4	846	48	1	710
21	3	1036	49	1	387
22	4	948	50	10	3516
23	4	1412	51	5	2002
24	3	438	52	9	3622
25	5	2111	53	2	1400
26	2	977	54	2	1584
27	3	814	55	3	830
28	1	319	56	8	167

TABLE 6.12—*Contd.*

Circle No. (i)	Number of Villages (M_i)	Area under Wheat (Acres) $(M_i\bar{y}_{i.})$	Circle No. (i)	Number of Villages (M_i)	Area under Wheat (Acres) $(M_i\bar{y}_{i.})$
57	3	622	75	4	669
58	2	591	76	1	1187
59	5	273	77	2	852
60	2	781	78	1	51
61	2	1101	79	1	1265
62	2	799	80	8	1423
63	2	601	81	2	794
64	3	928	82	1	1604
65	4	1141	83	3	1621
66	1	1208	84	2	1764
67	5	1633	85	6	2668
68	4	902	86	1	1076
69	3	1286	87	1	348
70	5	1299	88	4	1224
71	7	1947	89	4	1490
72	3	741			
73	2	574	Total	299	98078
74	7	2554			

TABLE 6.13

Analysis of Variance of Areas under Wheat in Villages in Hapur Subdivision (Acres)^2

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square
Between circles ..	88	10924581	124143
Within circles between villages ..	210	9588011	45657
Total population ..	298	20512592	68834

It will be noticed that the bias exceeds the standard error proper owing to the large variation in M_i. The method must, therefore, be rejected from further consideration.

(b) Equal Probability, Mean of the Cluster Totals Estimate

$$V\left(\frac{N}{n}\sum^{n} M_i\bar{y}_{i.}\right)$$

$$= N^2\left\{\frac{N-n}{Nn}\cdot\frac{1}{N-1}\sum_{i=1}^{N}\left(M_i\bar{y}_{i.}-\frac{1}{N}\sum_{i=1}^{N}M_i\bar{y}_{i.}\right)^2\right\}$$

$$= 89^2\cdot\frac{69}{1780}\cdot 513613$$

$$= 1577\times 10^5 \text{ acres}^2$$

(c) Equal Probability, Ratio Estimate

$$V_1\left\{\frac{M_0\overset{n}{\Sigma} M_i\bar{y}_{i.}}{\overset{n}{\Sigma} M_i}\right\}$$

$$= M_0^2\left\{\frac{N-n}{Nn}\cdot\frac{1}{\bar{M}^2}\cdot\frac{1}{N-1}\sum_{i=1}^{N}M_i^2\,(\bar{y}_{i.}-\bar{y}_{..})^2\right\}$$

$$= \frac{69\times 89}{20}\cdot 342043$$

$$= 1050\times 10^5 \text{ acres}^2$$

$$V_2\left\{\frac{M_0\overset{n}{\Sigma} M_i\bar{y}_{i.}}{\overset{n}{\Sigma} M_i}\right\} = V_1\left(1+\frac{3}{n}\,(C.V.\ M_i)^2\right)$$

$$= 1050\left(1+\frac{3}{20}\cdot\frac{4{\cdot}074}{(3{\cdot}360)^2}\right)\times 10^5$$

$$= 1107\times 10^5 \text{ acres}^2$$

(*d*) *Probability Proportional to the Size of the Circle, Mean of the Cluster Means Estimate*

$$V(M_0\bar{y}_{n.}) = M_0^2 \cdot \frac{1}{n} \sum_{i=1}^{N} \frac{M_i}{M_0} (\bar{y}_{i.} - \bar{y}_{..})^2$$

$$= (299)^2 \cdot \frac{1}{20} \cdot (36537)$$

$$= 1633 \times 10^5 \text{ acres}^2$$

(*e*) *Village as Unit of Sampling, Equal Probability, Mean per Village Estimate*

$$V(M_0\bar{y}_{n\bar{M}}) = M_0^2 \cdot \frac{N\bar{M} - n\bar{M}}{N\bar{M}} \cdot \frac{1}{n\bar{M}} \cdot \frac{1}{M_0 - 1} \sum_{i=1}^{M_0} (y_{ij} - \bar{y}_{..})^2$$

$$= \frac{299 \times 69}{20} (68834)$$

$$= 710 \times 10^5 \text{ acres}^2$$

The relative efficiencies of the different methods are then as follows:

Sampling Unit	Sampling Method	Method of Estimation	Relative Efficiency
(*a*) Circle	Equal probability	Mean of cluster means	12
(*b*) Circle	Equal probability	Mean of cluster totals	45
(*c*) Circle	Equal probability	Ratio	64
(*d*) Circle	Probability proportional to size	Mean of cluster means	43
(*e*) Village	Equal probability	Mean per village	100

The very low efficiency of method (*a*) is partly due to the presence of serious bias in the estimate. Of the other methods with the circle as the unit of sampling, the ratio method is seen to be the most efficient. The explanation is provided by Table 6.14 showing the two-way classification of circles by the area under wheat and by size in terms of the number of villages.

TABLE 6.14

Frequency Distribution of Circles by Area under Wheat and Number of Villages

Area under Wheat in a Circle										
3600<3800									1	
<3600										1
<3400										
<3200										
<3000										
<2800						1	1			
<2600						1	1	1		
<2400										
<2200				2	2					
<2000							1			
<1800	1	1	1	1	1					
<1600		3	1	2		1		1		
<1400	2		1	3	1					
<1200	2	1	2	3	1					
<1000		3	5	3						
< 800	2	7	4	2	1					
< 600	1	7	1	1	1					
200< 400	5				2					
0< 200	2							1		
	1	2	3	4	5	6	7	8	9	10

Number of Villages in a Circle

The relationship between the two, *i.e*, $M_i\bar{y}_{i.}$ and M_i, is seen to be approximately linear, the value of the coefficient of correlation being 0·64. The variability among areas under wheat in circles of the same size is seen to be rather independent of the size, and explains the relative superiority of the ratio method (with equal probability) over the simple arithmetic mean estimate when the clusters are selected with probability proportional to size. The village as a unit of sampling is seen to be far superior to the circle, and since the village is also known to be administratively convenient, it is preferred to the use of the circle as the unit of sampling in most agricultural surveys in India.

REFERENCES

1. Hansen, M. H. and Hurwitz, W. N. (1942) "Relative Efficiencies of Various Sampling Units in Population Inquiries," *Jour. Amer. Statist. Assoc.*, **37**, 89–94.
2. Smith, H. Fairfield (1938) "An Empirical Law Describing Heterogeneity in the Yields of Agricultural Crops," *Jour. Agr. Sci.*, **28**, 1–23.
3. Jessen, R. J. (1942) .. "Statistical Investigation of a Sample Survey for Obtaining Farm Facts," *Iowa Agricultural Experiment Station, Research Bulletin* 304.
4. Mahalanobis, P. C. (1940) "A Sample Survey of the Acreage under Jute in Bengal," *Sankhya*, **4**, 511–30.
5. Hendricks, W. A. (1944) .. "The Relative Efficiencies of Groups of Farms as Sampling Units," *Jour. Amer. Statist. Assoc.*, **39**, 366–76.
6. Asthana, R. S. (1950) .. "The Size of Sub-Sampling Unit in Area Estimation," Unpublished Thesis for Diploma, I.C.A.R., New Delhi.
7. Sukhatme, P. V. (1950) .. "Sample Surveys in Agriculture," *Presidential Address to the Section of Statistics, 37th Session, Indian Science Congress, Poona.*
8. —— and Panse, V. G. (1951) "Crop Surveys in India—II," *Jour. Ind. Soc. Agr. Statist.*, **3**, 97–168.
9. Cochran, W. G. (1948) .. *Notes on 'Sample Survey Techniques'* (Mimeographed), Institute of Statistics, Raleigh, North Carolina.

CHAPTER VII

SUB-SAMPLING

7.1 Introduction

So far we have considered only sampling procedures in which all the elements of the selected clusters are enumerated. We also saw that the larger the cluster the less efficient it will usually be relative to the element as the unit of sampling. It is therefore logical to expect that, for a given number of elements, greater precision will be attained by distributing them over a large number of clusters than by taking a small number of clusters and sampling a large number of elements from each of them or completely enumerating them. The procedure of first selecting clusters and then choosing a specified number of elements from each selected cluster is known as *sub-sampling*. It is also known as *two-stage sampling*. The clusters which form the units of sampling at the first stage are called the *first-stage units* and the elements or groups of elements within clusters which form the units of sampling at the second stage are called *sub-units* or *second-stage units*. The procedure is easily generalized to three-stage or *multi-stage sampling*. As an example of three-stage sampling, we may refer to crop surveys in which villages are the first-stage units, fields within villages are the second-stage units and plots within fields the third-stage units of sampling, the correlation between yield of adjoining portions of the same field rendering it unnecessary and also uneconomical to harvest the whole field.

7.2 Two-Stage Sampling, Equal First-Stage Units: Estimate of the Population Mean

We shall first consider the case of equal clusters and assume that the population is composed of NM elements grouped into N first-stage units of M second-stage units each. Let n denote the number of first-stage units in the sample and m the number of second-stage units to be drawn from each selected first-stage unit. Further, we shall suppose that the units at each stage are selected with equal probability.

Now let, as previously,

y_{ij} = the value of the j-th second-stage unit in the i-th first-stage unit ($j = 1, 2, \ldots, M$; $i = 1, 2, \ldots, N$)

$\bar{y}_{i.}$ = the mean per second-stage unit in the i-th first-stage unit in the population ($i = 1, 2, \ldots, N$)

and

$$\bar{y}_{..} = \frac{1}{NM} \sum_{i=1}^{N} \sum_{j=1}^{M} y_{ij}$$

= the mean per second-stage unit in the population

Further, let

$$\bar{y}_{im} = \frac{1}{m} \sum_{j}^{m} y_{ij}$$

= the mean per second-stage unit of the i-th first-stage unit in the sample

and

$$\bar{y}_{nm} = \frac{1}{nm} \sum_{i}^{n} \sum_{j}^{m} y_{ij}$$

$$= \frac{1}{n} \sum^{n} \bar{y}_{im}$$

= the mean per second-stage unit in the sample

Then it can be shown that of all the linear estimates of the type $\sum_{i}^{n} \sum_{j}^{m} d_{ij}y_{ij}$, where the d_{ij} are constants independent of the selected sample, the sample mean $\bar{y}_{nm}$ provides the best unbiased estimate of the population mean $\bar{y}_{..}$. In the next section we shall derive the expected value and the sampling variance of this estimate but the proof that it is the best linear estimate is left to the reader.

7.3 Two-Stage Sampling, Equal First-Stage Units: Expected Value and Variance of the Sample Mean

Since the sample is selected in two stages, the expected value is also appropriately worked out in two stages, first over all possible samples of m from each of n fixed first-stage units and then over all possible samples of n. Thus, we have

$$\begin{aligned} E(\bar{y}_{nm}) &= E\left\{\frac{1}{n}\sum^{n}\bar{y}_{im}\right\} \\ &= E\left\{\frac{1}{n}\sum^{n}E(\bar{y}_{im} \mid i)\right\} \\ &= E\left\{\frac{1}{n}\sum^{n}\bar{y}_{i.}\right\} \\ &= \bar{y}_{N.} \end{aligned}$$

Since the first-stage units are equal,

$$\bar{y}_{N.} = \bar{y}_{..}$$

Hence

$$E(\bar{y}_{nm}) = \bar{y}_{..} \tag{1}$$

thus showing that the simple mean of all elements in the sample gives an unbiased estimate of the population mean.

By definition, the variance of the sample mean is given by

$$\begin{aligned} V(\bar{y}_{nm}) &= E\{\bar{y}_{nm} - E(\bar{y}_{nm})\}^2 \\ &= E(\bar{y}_{nm} - \bar{y}_{N.})^2 \end{aligned} \tag{2}$$

This can be written as

$$\begin{aligned} V(\bar{y}_{nm}) &= E(\bar{y}_{nm} - \bar{y}_{n.} + \bar{y}_{n.} - \bar{y}_{N.})^2 \\ &= E(\bar{y}_{nm} - \bar{y}_{n.})^2 + E(\bar{y}_{n.} - \bar{y}_{N.})^2 \\ &\qquad + 2E\{(\bar{y}_{nm} - \bar{y}_{n.})(\bar{y}_{n.} - \bar{y}_{N.})\} \end{aligned} \tag{3}$$

where $\bar{y}_{n.}$ denotes the simple mean of n first-stage unit means, given by

$$\bar{y}_{n.} = \frac{1}{n} \sum^{n} \bar{y}_{i.}$$

Now

$$\bar{y}_{nm} - \bar{y}_{n.} = \frac{1}{n} \sum^{n} (\bar{y}_{im} - \bar{y}_{i.})$$

so that

$$\begin{aligned} E\,(\bar{y}_{nm} - \bar{y}_{n.})^2 &= \frac{1}{n^2}\, E\left[\sum^{n} (\bar{y}_{im} - \bar{y}_{i.})\right]^2 \\ &= \frac{1}{n^2}\, E\left[\sum^{n} (\bar{y}_{im} - \bar{y}_{i.})^2 + \sum_{i \neq i'}^{n} (\bar{y}_{im} - \bar{y}_{i.})\,(\bar{y}_{i'm} - \bar{y}_{i'.})\right] \\ &= \frac{1}{n^2}\left[E \sum^{n} E\,\{(\bar{y}_{im} - \bar{y}_{i.})^2 \mid i\}\right. \\ &\qquad \left. + E \sum_{i \neq i'}^{n} E\,\{(\bar{y}_{im} - \bar{y}_{i.})\,(\bar{y}_{i'm} - \bar{y}_{i'.}) \mid i, i'\}\right] \quad (4) \end{aligned}$$

The value of the second term under the summation sign is clearly zero since sub-samples are drawn independently from the i-th and i'-th first-stage units and the value of the first term under the summation sign is given by the well-known result

$$E\,\{(\bar{y}_{im} - \bar{y}_{i.})^2 \mid i\} = \left(\frac{1}{m} - \frac{1}{M}\right) \mathrm{S}_i^2 \qquad (5)$$

where

$$\mathrm{S}_i^2 = \frac{\sum_{j=1}^{M} (y_{ij} - \bar{y}_{i.})^2}{M - 1}$$

whence we obtain

$$\begin{aligned} E\,(\bar{y}_{nm} - \bar{y}_{n.})^2 &= \frac{1}{n^2}\, E \sum^{n} \left(\frac{1}{m} - \frac{1}{M}\right) \mathrm{S}_i^2 \\ &= \frac{1}{nN} \sum_{i=1}^{N} \left(\frac{1}{m} - \frac{1}{M}\right) \mathrm{S}_i^2 \\ &= \frac{1}{n} \left(\frac{1}{m} - \frac{1}{M}\right) \bar{\mathrm{S}}_w^2 \qquad (6) \end{aligned}$$

where

$$\bar{S}_w^2 = \frac{1}{N} \sum_{i=1}^{N} S_i^2 \tag{7}$$

Also, from (5) and (6) of Chapter VI,

$$E\,(\bar{y}_{n.} - \bar{y}_{N.})^2 = \left(\frac{1}{n} - \frac{1}{N}\right) S_b^2 \tag{8}$$

where

$$S_b^2 = \frac{\sum\limits_{i=1}^{N} (\bar{y}_{i.} - \bar{y}_{N.})^2}{N-1}$$

The value of the last term in (3) is clearly zero. For,

$$\begin{aligned} & E\,\{(\bar{y}_{nm} - \bar{y}_{n.})\,(\bar{y}_{n.} - \bar{y}_{N.})\} \\ &\qquad = E\left[(\bar{y}_{n.} - \bar{y}_{N.}) \times \frac{1}{n} \sum^{n} E\,\{(\bar{y}_{im} - \bar{y}_{i.}) \mid i\}\right] \\ &\qquad = E\,[(\bar{y}_{n.} - \bar{y}_{N.}) \times 0] \\ &\qquad = 0 \end{aligned} \tag{9}$$

Substituting from (6), (8) and (9) in (3), we get on interchanging the order of the first two terms

$$V(\bar{y}_{nm}) = \left(\frac{1}{n} - \frac{1}{N}\right) S_b^2 + \left(\frac{1}{m} - \frac{1}{M}\right) \frac{\bar{S}_w^2}{n} \tag{10}$$

The variance is seen to be made up of two components. If the selected first-stage units had been completely enumerated, in other words, if $m = M$, the variance of the sample mean would be obviously given by the first component only. Actually, each selected first-stage unit is only partially enumerated by means of a sample of second-stage units drawn from it. The second term in (10) thus represents the contribution to the sampling variance arising from sub-sampling the selected first-stage units. In fact, setting $m = M$ in (10) we have equation (5) of Chapter VI.

When $n = N$, or in other words, every first-stage unit in the population is sampled, we are left with the second component to represent the variance of the sample mean. This case corresponds to stratified sampling with first-stage units as strata and a simple random sample of m drawn from each of several strata. We can thus look upon a sub-sampling design as a case of incomplete stratification as it were, the first component representing the additional contribution to the variance of a stratified sample arising from incomplete stratification.

When N is large relative to n, so that $(N - n)/N$ can be taken as unity, we have

$$V(\bar{y}_{nm}) = \frac{\mathrm{S}_b^2}{n} + \left(\frac{1}{m} - \frac{1}{M}\right)\frac{\bar{\mathrm{S}}_w^2}{n} \qquad (11)$$

When M is large relative to m, so that $(M - m)/M$ can be taken as unity, we have

$$V(\bar{y}_{nm}) = \left(\frac{1}{n} - \frac{1}{N}\right)\mathrm{S}_b^2 + \frac{\bar{\mathrm{S}}_w^2}{nm} \qquad (12)$$

And when finite multipliers at both stages can be taken as unity, we are left with the simple expression

$$V(\bar{y}_{nm}) = \frac{\mathrm{S}_b^2}{n} + \frac{\bar{\mathrm{S}}_w^2}{nm} \qquad (13)$$

7.4 Two-Stage Sampling, Equal First-Stage Units: Estimation of the Variance of the Sample Mean

The calculation of the variance of the sample mean in two-stage sampling involves the estimation of S_b^2 and $\bar{\mathrm{S}}_w^2$. The simplest way of estimating these is to define the corresponding quantities for the sample and obtain their expected values.

Let s_b^2 denote the mean square between first-stage unit means in the sample defined by

$$s_b^2 = \frac{\sum^{n} (\bar{y}_{im} - \bar{y}_{nm})^2}{n - 1} \qquad (14)$$

and s_i^2 denote the mean square between second-stage units drawn from the i-th first-stage unit defined by

$$s_i^2 = \frac{\sum_j^m (y_{ij} - \bar{y}_{im})^2}{m - 1} \tag{15}$$

Equation (14) can be rewritten as

$$(n - 1)\, s_b^2 = \sum^n \bar{y}_{im}^2 - n\bar{y}_{nm}^2$$

whence

$$(n - 1)\, E\,(s_b^2) = E\left(\sum^n \bar{y}_{im}^2\right) - nE\,(\bar{y}_{nm}^2) \tag{16}$$

Now, to evaluate the first term in (16), we write

$$\begin{aligned} E\left(\sum^n \bar{y}_{im}^2\right) &= E\left\{\sum^n E\,(\bar{y}_{im}^2 \mid i)\right\} \\ &= E\left[\sum^n \left\{\bar{y}_{i.}^2 + \left(\frac{1}{m} - \frac{1}{M}\right) \mathrm{S}_i^2\right\}\right] \\ &= \frac{n}{N}\left[\sum_{i=1}^N \bar{y}_{i.}^2 + N\left(\frac{1}{m} - \frac{1}{M}\right) \bar{\mathrm{S}}_w^2\right] \end{aligned} \tag{17}$$

The value of the second term in (16) can be directly obtained from (10). For, by definition,

$$V\,(\bar{y}_{nm}) = E\,(\bar{y}_{nm}^2) - \bar{y}_{N.}^2$$

whence

$$nE\,(\bar{y}_{nm}^2) = \left(1 - \frac{n}{N}\right) \mathrm{S}_b^2 + \left(\frac{1}{m} - \frac{1}{M}\right) \bar{\mathrm{S}}_w^2 + n\bar{y}_{N.}^2 \tag{18}$$

Substituting from (17) and (18) in (16), and recalling that

$$(N - 1)\, \mathrm{S}_b^2 = \sum_{i=1}^N \bar{y}_{i.}^2 - N\bar{y}_{N.}^2$$

we obtain

$$E\,(s_b^2) = \mathrm{S}_b^2 + \left(\frac{1}{m} - \frac{1}{M}\right) \bar{\mathrm{S}}_w^2 \tag{19}$$

From (36) of Chapter II, we know that for fixed i,

$$E\,(s_i^2) = \mathbf{S}_i^2$$

Also, for varying i,

$$E\left(\frac{1}{n}\sum^{n}\mathbf{S}_i^2\right) = \frac{1}{N}\sum_{i=1}^{N}\mathbf{S}_i^2 = \bar{\mathbf{S}}_w^2$$

whence

$$E\left(\frac{1}{n}\sum^{n} s_i^2\right) = \bar{\mathbf{S}}_w^2 \tag{20}$$

We thus have

$$\text{Est. } \bar{\mathbf{S}}_w^2 = \bar{s}_w^2$$

$$= \frac{\sum^{n}(m-1)\,s_i^2}{n\,(m-1)} \tag{21}$$

= mean square within first-stage units in the analysis of variance of the sample

and

$$\text{Est. } \mathbf{S}_b^2 = s_b^2 - \left(\frac{1}{m} - \frac{1}{M}\right)\bar{s}_w^2 \tag{22}$$

When $m = M$, the second term in (22) vanishes and we are left with the known result appropriate for one-stage sampling without sub-sampling. We note also that in two-stage sampling the estimate of $\mathbf{S}_b^2$ is less than the corresponding mean square between the first-stage unit means in the sample, as one would expect, since s_b^2 is based on estimates of first-stage unit means and not the true values and therefore subject to an additional component of variation.

Substituting from (21) and (22) in (10), we have

$$\text{Est. } V\,(\bar{y}_{nm}) = \left(\frac{1}{n} - \frac{1}{N}\right)s_b^2 + \frac{1}{N}\left(\frac{1}{m} - \frac{1}{M}\right)\bar{s}_w^2 \tag{23}$$

When $(N - n)/N$ can be taken as unity, (21) and (22) will still hold, giving on substitution in (11)

$$\text{Est. } V(\bar{y}_{nm}) = \frac{s_b^2}{n} \tag{24}$$

$$= \frac{\text{Mean square between first-stage units in the analysis of variance for the sample}}{nm}$$

When $(M - m)/M$ can be taken as unity, we have

$$\text{Est. } \bar{\mathbf{S}}_w^2 = \bar{s}_w^2$$

$$\text{Est. } \mathbf{S}_b^2 = s_b^2 - \frac{\bar{s}_w^2}{m} \tag{25}$$

and

$$\text{Est. } V(\bar{y}_{nm}) = \left(\frac{1}{n} - \frac{1}{N}\right) s_b^2 + \frac{1}{Nm}\, \bar{s}_w^2 \tag{26}$$

When $(N - n)/N$ and $(M - m)/M$ can each be taken as unity, (25) holds good, giving on substitution in (13), for an estimate of $V(\bar{y}_{nm})$, an expression identical with (24), namely,

$$\text{Est. } V(\bar{y}_{nm}) = \frac{s_b^2}{n} \tag{27}$$

$$= \frac{\text{Mean square between first-stage units in the analysis of variance of the sample}}{nm}$$

7.5 Distribution of Sample between Two Stages: Equal First-Stage Units

The expression (10) for the variance of the sample mean in a two-stage sampling design shows that the precision of a two-stage sample, apart from the values of $\mathbf{S}_b^2$ and $\bar{\mathbf{S}}_w^2$, depends upon the distribution of the sample between the two stages, or in other words, on n and m individually. The cost of surveying a two-stage sample will likewise depend upon the values of n and m. In this section we shall consider the problem of choosing n and m so that the variance of the sample mean is minimized for given cost. Alternatively, we can choose n and m so as to provide an estimate of the desired precision for minimum cost.

We shall first consider the simplest case in which the cost of the survey is proportional to the size of the sample, so that

$$C = cnm \tag{28}$$

where C = the total cost of the survey and c is a constant. Let the total cost of the survey be fixed at, say, $C = C_0$. Then from (28), we have

$$m = \frac{C_0}{cn} \tag{29}$$

Substituting from (29) in the expression for the variance of the sample mean given by (10), we obtain

$$\begin{aligned} V(\bar{y}_{nm}) &= \left(\frac{1}{n} - \frac{1}{N}\right) \mathrm{S}_b^2 + \left(\frac{c}{C_0} - \frac{1}{Mn}\right) \bar{\mathrm{S}}_w^2 \\ &= \left(\mathrm{S}_b^2 - \frac{\bar{\mathrm{S}}_w^2}{M}\right)\frac{1}{n} - \frac{\mathrm{S}_b^2}{N} + \frac{c\bar{\mathrm{S}}_w^2}{C_0} \end{aligned} \tag{30}$$

which is a monotonic decreasing function of n if $\mathrm{S}_b^2 - \bar{\mathrm{S}}_w^2/M$ is positive, reaching its minimum when n assumes the maximum value, namely,

$$\hat{n} = \frac{C_0}{c} \tag{31}$$

Equation (29) then gives

$$\hat{m} = 1 \tag{32}$$

If $\mathrm{S}_b^2 - \bar{\mathrm{S}}_w^2/M < 0$, which for large N is equivalent to stating that the intra-class correlation is negative, the right-hand side of equation (30) becomes a monotonic increasing function of n, reaching its minimum when n is minimum, given by

$$\hat{n} = \frac{C_0}{cM}$$

In other words, there is no sub-sampling.

The alternative approach of estimating the population mean with the desired precision for minimum cost leads to the same solution for m. For, let V_0 be the value of the variance with which it is desired to estimate the population mean, so that

$$V_0 = \left(\frac{1}{n} - \frac{1}{N}\right) \mathrm{S}_b^2 + \left(\frac{1}{m} - \frac{1}{M}\right) \frac{\bar{\mathrm{S}}_w^2}{n} \tag{33}$$

Solving for n, we get

$$n = \frac{\mathrm{S}_b^2 + \left(\frac{1}{m} - \frac{1}{M}\right) \bar{\mathrm{S}}_w^2}{V_0 + \frac{\mathrm{S}_b^2}{N}} \tag{34}$$

On substituting in (28), we obtain

$$C = cm \left\{ \frac{\mathrm{S}_b^2 - \frac{\bar{\mathrm{S}}_w^2}{M}}{V_0 + \frac{\mathrm{S}_b^2}{N}} \right\} + \frac{c\bar{\mathrm{S}}_w^2}{V_0 + \frac{\mathrm{S}_b^2}{N}} \tag{35}$$

Clearly, for $\mathrm{S}_b^2 - \bar{\mathrm{S}}_w^2/M > 0$, C attains its minimum when m assumes the smallest integral value, namely 1; and for $\mathrm{S}_b^2 - \bar{\mathrm{S}}_w^2/M < 0$, the minimum is attained when $\hat{m} = M$.

It should be noted that the optimum distribution is independent of the variability of the character under study. This suggests the advisability of enlarging the scope of the questionnaire to carry several items whenever the field cost is proportional to the number of second-stage units or interviews, the sub-sampling design with one sub-unit from each selected first-stage unit being the most efficient in this case, provided, of course, $\mathrm{S}_b^2 - \bar{\mathrm{S}}_w^2/M$ is positive for each item, and $C_0/c < N$.

We shall next consider a more general case when the cost of the survey is represented by

$$C = c_1 n + c_2 nm \tag{36}$$

where c_1 and c_2 are positive constants. From (36) and (10), we obtain

$$C\left\{V(\bar{y}_{nm}) + \frac{S_b^2}{N}\right\} = (c_1 + c_2 m)\left[\left(S_b^2 - \frac{1}{M}\bar{S}_w^2\right) + \frac{1}{m}\bar{S}_w^2\right]$$

$$= c_1\left(S_b^2 - \frac{1}{M}\bar{S}_w^2\right) + c_2\bar{S}_w^2$$

$$+ mc_2\left(S_b^2 - \frac{1}{M}\bar{S}_w^2\right) + \frac{c_1\bar{S}_w^2}{m} \qquad (37)$$

Clearly, the minimum value of (37) will provide the optimum allocation for both the cases, when either C or V is fixed in advance and V or C minimized.

For $S_b^2 - \bar{S}_w^2/M > 0$, (37) can be put in the form

$$C\left(V + \frac{1}{N}S_b^2\right) = \left[\sqrt{c_1\left(S_b^2 - \frac{1}{M}\bar{S}_w^2\right)} + \sqrt{c_2\bar{S}_w^2}\right]^2$$

$$+ \left[\sqrt{mc_2\left(S_b^2 - \frac{1}{M}\bar{S}_w^2\right)} - \sqrt{\frac{c_1\bar{S}_w^2}{m}}\right]^2$$

and is minimum when the square term in m is equated to zero, or in other words, when $\hat{m}$ is given by the nearest integer to

$$\sqrt{\frac{c_1}{c_2} \cdot \frac{\bar{S}_w^2}{S_b^2 - \frac{1}{M}\bar{S}_w^2}} \qquad (38)$$

or, approximately by

$$\sqrt{\frac{c_1}{c_2}\left(\frac{1}{\rho} - 1\right)} \qquad (39)$$

where ρ is the intra-class correlation within first-stage units.

For $S_b^2 - \bar{S}_w^2/M \leqslant 0$, the expression on the right-hand side of (37) is minimum when m is the maximum attainable integral value. If the total cost fixed for the survey, namely C_0, exceeds $c_1 + c_2 M$, we have

$$\hat{m} = M$$

and $\hat{n}$ is the greatest integer in

$$\frac{C_0}{c_1 + c_2 M} \tag{40}$$

If C_0 is less than $c_1 + c_2 M$, $\hat{m}$ is the greatest integer in

$$\frac{C_0 - c_1}{c_2}$$

and $\hat{n}$ is 1.

It will be noticed that $\hat{m}$ is now dependent upon the magnitude of the two cost constants, as also on the intra-class correlation of the character under study. In general, if $\mathbf{S}_b^2 - \bar{\mathbf{S}}_w^2/M > 0$, the optimum value for m will be smaller if: (i) the travel between first-stage units and other costs which go to make up c_1 are cheaper, (ii) the cost of collecting sub-samples from the selected first-stage units is larger, and (iii) the intra-class correlation is large. It follows that the optimum sub-sampling rate will change from character to character. For a related group of items, however, for which the value of ρ does not materially vary it may be possible to hit upon a satisfactory optimum for m without appreciable loss of efficiency. Sample surveys for the estimation of acreage under major crops provide evidence on this point. Table 7.1 gives the values of $\mathbf{S}_b^2$, $\bar{\mathbf{S}}_w^2$ and ρ for the acreage under wheat, gram, maize and sugarcane calculated from the records of complete enumeration during 1936 of all the villages in Hapur Subdivision of Meerut District (India). The village constituted the first-stage unit and a grid of 8 fields formed by grouping consecutive fields within a village was taken as the second-stage unit. It is seen that ρ is of the same order for three out of the four crops and it seems feasible to hit upon a satisfactory value of m without very much departing from the optimum for individual crops. When, however, minor crops are also included in the survey, the value of ρ is found to vary considerably and it is not possible to design an efficient two-stage sample with a common sub-sampling rate for all items. This is true of all general purpose surveys and one solution at least of minimizing the loss of efficiency lies in grouping together related items of the questionnaire and using different sampling designs for the several groups.

TABLE 7.1

Values of the True Variance between and within Villages for Area under Different Crops in Hapur Subdivision (Biswas^2/Grid)

	Area under Crops in 1000 Biswas*	S_b^2	$\bar{S}_w^2$	ρ
Wheat	1962	483·8	4991·4	·0884
Gram	552	121·2	700·2	·1476
Maize	556	60·3	698·2	·0795
Sugarcane	526	110·6	1127·7	·0893

* Biswas is a local unit of area.

Example 7.1

A yield survey on paddy was carried out in West Godavari District (India) in 1946–47. Five villages were selected in each of the seven strata into which the district is divided, three fields were harvested in each village and one plot of 1/100 acre was harvested in each field. The data are reproduced in Table 7.2. Obtain pooled values of s_b^2 and $\bar{s}_w^2$ for the district and the estimates of S_b^2 and $\bar{S}_w^2$. Finite multipliers at the sub-sampling stage may be ignored.

Calculate the sampling variance of the estimate of the district mean yield and the percentage standard error.

Assuming that the sample of villages is to be allocated in proportion to the numbers of villages in the several strata and that the cost in rupees of the survey is represented by

$$C = 7n + 2nm$$

calculate the values of n and m that may be recommended for a subsequent survey in order that the district mean yield may be estimated with standard error of 12 ozs. per plot for the minimum cost.

From the last row of Table 7.2, we obtain

$$\text{Pooled } s_b^2 = \frac{\sum_{t=1}^{k} (n_t - 1)\, s_{tb}^2}{n - k}$$

$$= \frac{4 \times (46655 \cdot 2)}{28}$$

$$= 6665 \cdot 0$$

$$\text{Pooled } \bar{s}_w^2 = \frac{\sum_{t=1}^{k} n_t\, (m-1)\, s_{tw}^2}{n\,(m-1)}$$

$$= \frac{66979 \cdot 5}{7}$$

$$= 9568 \cdot 5$$

On substituting in (25), we obtain

$$\text{Est. } \mathrm{S}_b^2 = 6665 \cdot 0 - \frac{9568 \cdot 5}{3}$$

$$= 3475 \cdot 5$$

and

$$\text{Est. } \bar{\mathrm{S}}_w^2 = 9568 \cdot 5$$

The variance of the district mean when finite multipliers at the sub-sampling stage are ignored is given by

$$V(\bar{y}_{nm}) = \mathrm{S}_b^2 \sum_{t=1}^{k} \left(\frac{1}{n_t} - \frac{1}{N_t}\right) \frac{N_t^2}{N^2} + \frac{\bar{\mathrm{S}}_w^2}{m} \sum_{t=1}^{k} \frac{1}{n_t}\, \frac{N_t^2}{N^2}$$

$$= \mathrm{S}_b^2 \left\{\frac{\sum_{t=1}^{k} p_t^2}{5} - \frac{1}{N}\right\} + \frac{\bar{\mathrm{S}}_w^2}{m} \cdot \frac{1}{5} \sum_{t=1}^{k} p_t^2$$

where $p_t = N_t/N$.

TABLE 7.2

Yield Survey on Paddy, West Godavari District (India), 1946–47

Values of Means, Mean Squares between Village Means (s_{tb}^2) and Mean Squares within Villages (s_{tw}^2) per Plot Basis

Stratum Number	No. of Villages in the		Sample Mean (Oz./Plot)	s_{tb}^2	s_{tw}^2	$\frac{N_t}{N}$	$\frac{N_t^2}{N^2}$
t	Population N_t	Sample n_t					
1	88	5	347·5	1452·1	2791·5	·109863	0·012070
2	142	5	297·8	1937·0	27422·5	·177278	0·031427
3	119	5	201·1	7107·2	1864·0	·148564	0·022071
4	90	5	438·9	9603·9	11824·0	·112360	0·012625
5	114	5	282·9	20702·0	13628·2	·142322	0·020256
6	102	5	301·9	2510·8	2007·5	·127341	0·016216
7	146	5	186·7	3342·2	7441·8	·182272	0·033223
Total	801	35		46655·2	66979·5		0·147888

Hence

$$\text{Est. } V(\bar{y}_{nm}) = 3475\cdot5 \left\{\frac{0\cdot147888}{5} - 0\cdot001248\right\} + \frac{9568\cdot5}{3} \times 0\cdot029578$$

$$= 192\cdot80$$

whence

$$S.E.\ (\bar{y}_{nm}) = 13\cdot89 \text{ oz./plot}$$

But

$$\bar{y}_{nm} = \sum_{t=1}^{k} p_t \bar{y}_{ntm}$$

$$= 282\cdot90 \text{ oz./plot}$$

whence

% *S.E.* of the estimate of district mean yield = 4·9

When the number of villages to be sampled is distributed between the various strata in proportion to N_t, the variance is given by

$$V(\bar{y}_{nm}) = \mathbf{S}_b^2\left(\frac{1}{n} - \frac{1}{N}\right) + \frac{\bar{\mathbf{S}}_w^2}{nm}$$

whence

$$n = \frac{\mathbf{S}_b^2 + \dfrac{\bar{\mathbf{S}}_w^2}{m}}{V(\bar{y}_{nm}) + \dfrac{\mathbf{S}_b^2}{N}}$$

For $V(\bar{y}_{nm}) = 144$ and $\mathbf{S}_b^2/N = 4\cdot3$, this reduces to

$$n = \frac{\mathbf{S}_b^2 + \dfrac{\bar{\mathbf{S}}_w^2}{m}}{148\cdot3}$$

Putting $m = 1, 2, 3, 4$ and 5 successively, we obtain the corresponding values of n. Substituting these in the equation for cost, we get the corresponding values of cost. The relevant calculations are given in Table 7.3.

TABLE 7.3

Calculations of the Optimum Sub-Sampling Rate

m	$\dfrac{\bar{\mathbf{S}}_w^2}{m}$	$\mathbf{S}_b^2 + \dfrac{\bar{\mathbf{S}}_w^2}{m}$	$n = \dfrac{(3)}{148\cdot3}$	C
(1)	(2)	(3)	(4)	(5)
1	9568·5	13044·0	88	792
2	4784·2	8259·7	56	616
3	3189·5	6665·0	45	585
4	2392·1	5867·6	40	600
5	1913·7	5389·2	36	612

It is seen that the cost is minimum when $m = 3$ and $n = 45$.

Alternatively, we can substitute directly in the equation for m, namely,

$$m^2 = \frac{c_1}{c_2} \times \frac{\bar{\mathrm{S}}_w^2}{\mathrm{S}_b^2}$$

which to the nearest integer gives

$$m = 3$$

The corresponding value for n is obtained by substitution in the equation for the variance and is found to be 45.

7.6 Comparison of Two-Stage with One-Stage Sampling

One-stage sampling procedures comparable with two-stage sampling will involve either

(i) sampling nm elements in one single stage, or

(ii) sampling of nm/M first-stage units as clusters, without further sub-sampling.

The variance of the mean of a simple random sample of nm elements selected by procedure (i) is given by

$$\left(\frac{1}{nm} - \frac{1}{NM}\right) \mathrm{S}^2 \tag{41}$$

To examine how this compares with the variance of a two-stage sample, it is convenient to express the latter in terms of the intra-class correlation between elements of the first-stage units. Substituting for S_b^2 and $\bar{\mathrm{S}}_w^2$ from (20) and (21) of the previous chapter in (10), we obtain

$$\begin{aligned} V(\bar{y}_{nm})_{Two\text{-}stage} &= \frac{NM-1}{NM} \cdot \frac{\mathrm{S}^2}{nm} \left[\left(1 - \frac{m}{M}\right)(1-\rho) \right. \\ &\qquad \left. + \frac{N-n}{N-1}\,\frac{m}{M}\,\{1 + (M-1)\,\rho\}\right] \\ &= \frac{NM-1}{NM} \cdot \frac{\mathrm{S}^2}{nm}\left[1 - \frac{m\,(n-1)}{M\,(N-1)} \right. \\ &\qquad + \rho\,\left\{\frac{N-n}{N-1} \cdot \frac{m}{M}\,(M-1) \right. \\ &\qquad \left.\left. - \frac{M-m}{M}\right\}\right] \end{aligned} \tag{42}$$

When the sub-sampling rate m/M is small, (42) may be approximated by

$$V(\bar{y}_{nm})_{Two\text{-}stage} \cong \frac{\mathbf{S}^2}{nm}\left[1+\rho\left(\frac{N-n}{N-1}\,m-1\right)\right] \tag{43}$$

Comparing this with (41), we notice that the relative change in variance using sub-sampling in place of unrestricted sampling of elements is approximately given by

$$\rho\left(\frac{N-n}{N-1}\,m-1\right) \tag{44}$$

This means that the relative efficiency of sub-sampling compared to unrestricted sampling of elements is approximately equal to that of sampling clusters of size $m\,(N-n)/(N-1)$. For n small compared with N, the relative change in variance is approximated by $\rho\,(m-1)$.

Next, the variance of the mean of an equivalent sample of nm/M clusters is given by

$$\left(\frac{M}{nm}-\frac{1}{N}\right)\mathbf{S}_b^2 \tag{45}$$

This exceeds (10) by

$$\frac{1}{n}\left(\frac{M}{m}-1\right)\left(\mathbf{S}_b^2-\frac{1}{M}\,\bar{\mathbf{S}}_w^2\right) \tag{46}$$

For N large and $\mathbf{S}_b^2 - \bar{\mathbf{S}}_w^2/M > 0$, the excess is approximated by

$$\frac{1}{n}\left(\frac{M}{m}-1\right)\rho\mathbf{S}^2 \tag{47}$$

showing that the smaller the sub-sampling rate m/M, the larger will be the reduction in variance of a two-stage sample over a cluster sample. When $\mathbf{S}_b^2 - \bar{\mathbf{S}}_w^2/M < 0$, sub-sampling will lead to loss of precision.

7.7 Effect of Change in Size of First-Stage Units on the Variance

We have seen in Section 7.6 that the variance of the mean of a two-stage sample consisting of n first-stage units with m second-stage units from each can be expressed as

$$V(\bar{y}_{nm}) = \frac{NM-1}{NM}\,\frac{\mathrm{S}^2}{nm}\left[1 - \frac{m(n-1)}{M(N-1)} + \rho_1\left\{\frac{N-n}{N-1}\,\frac{m}{M}(M-1) - \frac{M-m}{M}\right\}\right]$$

where ρ_1 represents the intra-class correlation within first-stage units of size M. We shall suppose now that the first-stage units are combined to give N/C new first-stage units with CM second-stage units each. The variance of the mean of a two-stage sample of size nm will then be given by

$$V'(\bar{y}_{nm}) \doteq \frac{NM-1}{NM}\,\frac{\mathrm{S}^2}{nm}\left[1 - \frac{m(n-1)}{M(N-C)} + \rho_2\left\{\frac{N-nC}{N-C}\,\frac{m}{MC}(MC-1) - \frac{MC-m}{MC}\right\}\right]$$

where ρ_2 will now represent the intra-class correlation within first-stage units of size MC. The difference between the two variances can be expressed as

$$V(\bar{y}_{nm}) - V'(\bar{y}_{nm}) = \frac{NM-1}{NM}\,\frac{\mathrm{S}^2}{nm}\left[\frac{m(n-1)(C-1)}{M(N-1)(N-C)} + \alpha_1\rho_1 - \alpha_2\rho_2\right]$$

where

$$\alpha_1 = \frac{N-n}{N-1}\,\frac{m}{M}(M-1) - \frac{M-m}{M}$$

$$\alpha_2 = \frac{N-nC}{N-C}\,\frac{m}{MC}(MC-1) - \frac{MC-m}{MC}$$

Since

$$\alpha_1 - \alpha_2 = \frac{m}{M}\left\{\frac{(C-1)(n-1)(NM-1)}{(N-1)(N-C)}\right\} \geqslant 0$$

and

$$\frac{m(n-1)(C-1)}{M(N-1)(N-C)} \geqslant 0$$

we conclude that

$$V(\bar{y}_{nm}) - V'(\bar{y}_{nm}) \geqslant 0$$

whenever $\rho_1 > \rho_2$ provided both ρ_1 and ρ_2 are positive. In other words, a gain in precision is brought about by enlarging first-stage units whenever the intra-class correlation is positive and decreases as the size of the first-stage unit increases. It also follows that the smaller the value of ρ_2 the larger is the gain, so that by choosing for consolidation those first-stage units which are as different as possible the gain can be increased. Practical considerations, however, put a limit on the size to which the first-stage units can be increased since cost of sub-sampling increases with larger and larger areas. This increase in precision is to be weighed against the increase in cost. As an example, we shall mention that in crop surveys, the variance is decreased when an administrative circle comprising a group of villages is used in place of a village as the first-stage unit of sampling, but practical considerations of cost and administrative convenience favour the use of the village (Sukhatme, 1950). If cost were no consideration, the enlargement of first-stage units could proceed to a point of eliminating the use of first-stage units altogether and the second-stage units would be selected independently from the whole population. This elegant analysis is due to Hansen and Hurwitz (1943).

7.8 Three-Stage Sampling, Equal First-Stage Units: Sample Mean and its Variance

Let

N = the number of first-stage units in the population

M = the number of second-stage units in each of N first-stage units

P = the number of third-stage units in each of NM second-stage units in the population

and n, m and p the corresponding values in the sample

Further, let

y_{ijk} = the value of the k-th element in the j-th second-stage unit of the i-th first-stage unit

$$\bar{y}_{ij.} = \frac{1}{P} \sum_{k=1}^{P} y_{ijk}$$

$$\bar{y}_{i..} = \frac{1}{MP} \sum_{j=1}^{M} \sum_{k=1}^{P} y_{ijk}$$

$$\bar{y}_{...} = \frac{1}{NMP} \sum_{i=1}^{N} \sum_{j=1}^{M} \sum_{k=1}^{P} y_{ijk}$$

and

$$\bar{y}_{ij(p)}, \quad \bar{y}_{i(mp)}, \quad \bar{y}_{(nmp)}$$

or, simply

$$\bar{y}_{ijp}, \quad \bar{y}_{imp}, \quad \bar{y}_{nmp}$$

denote the corresponding values for the sample, the use of parentheses to distinguish numbers on which the mean is based from the serial numbers of the selected units being made only where necessary.

We shall assume that the units at each stage are selected with equal probaoility.

It is easily shown that, as in two-stage sampling, the sample mean $\bar{y}_{nmp}$ provides an unbiased estimate of the population mean $\bar{y}_{...}$. For, we have

$$E(\bar{y}_{nmp}) = E\left\{\frac{1}{n} \sum^{n} E(\bar{y}_{imp} \mid i)\right\} \tag{48}$$

Since *mp* is a simple two-stage sample from the *i*-th first-stage unit, we have on substituting from (1) in (48),

$$\begin{aligned} E(\bar{y}_{nmp}) &= E\left\{\frac{1}{n} \sum^{n} \bar{y}_{i..}\right\} \\ &= \frac{1}{N} \sum_{i=1}^{N} \bar{y}_{i..} \\ &= \bar{y}_{...} \end{aligned} \tag{49}$$

To obtain the variance of $\bar{y}_{nmp}$, we write

$$
\begin{aligned}
V(\bar{y}_{nmp}) &= E(\bar{y}_{nmp} - \bar{y}_{...})^2 \\
&= E(\bar{y}_{nmp} - \bar{y}_{n..} + \bar{y}_{n..} - \bar{y}_{...})^2 \\
&= E\left[\left\{\frac{1}{n}\sum^{n}(\bar{y}_{imp} - \bar{y}_{i..})\right\}^2 + (\bar{y}_{n..} - \bar{y}_{...})^2\right. \\
&\qquad \left. + \frac{2}{n}\left\{\sum^{n}(\bar{y}_{imp} - \bar{y}_{i..})\right\}(\bar{y}_{n..} - \bar{y}_{...})\right] \\
&= E\left[\frac{1}{n^2}\sum^{n}(\bar{y}_{imp} - \bar{y}_{i..})^2 + \frac{1}{n^2}\sum_{i \neq i'}^{n}(\bar{y}_{imp} - \bar{y}_{i..})(\bar{y}_{i'mp} - \bar{y}_{i'..})\right. \\
&\qquad \left. + (\bar{y}_{n..} - \bar{y}_{...})^2 + \frac{2}{n}\left\{\sum^{n}(\bar{y}_{imp} - \bar{y}_{i..})\right\}(\bar{y}_{n..} - \bar{y}_{...})\right] \\
&= \frac{1}{n^2}E\left[\sum^{n}E\left\{(\bar{y}_{imp} - \bar{y}_{i..})^2 \mid i\right\}\right] \\
&\qquad + \frac{1}{n^2}E\left[\sum_{i \neq i'}^{n}E\left\{(\bar{y}_{imp} - \bar{y}_{i..})(\bar{y}_{i'mp} - \bar{y}_{i'..}) \mid i, i'\right\}\right] \\
&\qquad + E(\bar{y}_{n..} - \bar{y}_{...})^2 + \frac{2}{n}E\left[\left\{\sum^{n}E\{(\bar{y}_{imp} - \bar{y}_{i..}) \mid i\}\right\}\right. \\
&\qquad \left. \times (\bar{y}_{n..} - \bar{y}_{...})\right] \qquad (50)
\end{aligned}
$$

Since mp is a simple two-stage sample from the i-th first-stage unit, we have on substituting from (1) in the second and fourth terms and from (10) in the first term in (50), and noting that sampling from the *i*-th and *i′*-th first-stage units is carried out independently,

$$V(\bar{y}_{nmp}) = \frac{1}{n^2} E \sum^{n} \left\{ \left(\frac{1}{m} - \frac{1}{M} \right) \mathrm{S}_i^2 + \left(\frac{1}{p} - \frac{1}{P} \right) \frac{\bar{\mathrm{S}}_i^2}{m} \right\} + E(\bar{y}_{n..} - \bar{y}_{...})^2$$

$$= \frac{1}{nN} \sum_{i=1}^{N} \left(\frac{1}{m} - \frac{1}{M} \right) \mathrm{S}_i^2 + \frac{1}{nN} \sum_{i=1}^{N} \left(\frac{1}{p} - \frac{1}{P} \right) \frac{\bar{\mathrm{S}}_i^2}{m} + \left(\frac{1}{n} - \frac{1}{N} \right) \mathrm{S}_b^2$$

where

$$\mathrm{S}_i^2 = \frac{1}{M-1} \sum_{j=1}^{M} (\bar{y}_{ij.} - \bar{y}_{i..})^2$$

$$\bar{\mathrm{S}}_i^2 = \frac{1}{M} \sum_{j=1}^{M} \mathrm{S}_{ij}^2$$

$$= \frac{1}{M} \sum_{j=1}^{M} \frac{1}{P-1} \sum_{k=1}^{P} (y_{ijk} - \bar{y}_{ij.})^2$$

and

$$\mathrm{S}_b^2 = \frac{1}{N-1} \sum_{i=1}^{N} (\bar{y}_{i..} - \bar{y}_{...})^2$$

On interchanging terms and writing

$$\bar{\mathrm{S}}_w^2 = \frac{1}{N} \sum_{i=1}^{N} \mathrm{S}_i^2 \quad \text{and} \quad \bar{\bar{\mathrm{S}}}_p^2 = \frac{1}{N} \sum_{i=1}^{N} \bar{\mathrm{S}}_i^2$$

we have finally

$$V(\bar{y}_{nmp}) = \left(\frac{1}{n} - \frac{1}{N} \right) \mathrm{S}_b^2 + \left(\frac{1}{m} - \frac{1}{M} \right) \frac{\bar{\mathrm{S}}_w^2}{n} + \left(\frac{1}{p} - \frac{1}{P} \right) \frac{\bar{\bar{\mathrm{S}}}_p^2}{nm}, \qquad (51)$$

It will be seen that the variance of the sample mean is made up of three components corresponding to the three stages of sampling. If each of the nm selected second-stage units were completely enumerated, in other words if $p = P$, the variance would be given by the first two terms appropriate to the two-stage sampling design. If each of the n first-stage units in the sample were completely enumerated, in other words if $m = M$ and $p = P$, we will be left with the first term only representing the variance appropriate for one-stage sampling.

When $n = N$, and from each of N units a two-stage sample is drawn, we shall be left with the second and third terms to represent the variance of the mean. The case will correspond to a stratified two-stage sampling design with the first-stage units in the population constituting the strata.

Lastly, when finite multipliers are ignored, we get a simple expression for the variance, namely,

$$V(\bar{y}_{nmp}) = \frac{\mathbf{S}_b^2}{n} + \frac{\bar{\mathbf{S}}_w^2}{nm} + \frac{\bar{\bar{\mathbf{S}}}_p^2}{nmp} \tag{52}$$

7.9 Three-Stage Sampling, Equal First-Stage Units: Estimation from the Sample of the Variance of the Mean

As a first step we shall derive the expected values of the various mean squares in the sample.

Since $E(s_{ij}^2) = \mathbf{S}_{ij}^2$, we have from (20),

$$E(\bar{s}_i^2) = \bar{\mathbf{S}}_i^2$$

giving us

$$E(\bar{\bar{s}}_p^2) = \bar{\bar{\mathbf{S}}}_p^2 \tag{53}$$

Again, from (19), we have for a two-stage sample drawn from a specified first-stage unit,

$$E(s_i^2) = \mathbf{S}_i^2 + \left(\frac{1}{p} - \frac{1}{P}\right) \bar{\mathbf{S}}_i^2$$

whence

$$E(\bar{s}_w^2) = \bar{\mathbf{S}}_w^2 + \left(\frac{1}{p} - \frac{1}{P}\right) \bar{\bar{\mathbf{S}}}_p^2 \tag{54}$$

Finally

$$E\{(n-1)\ s_b^2\} = E\left\{\sum^{n} (\bar{y}_{imp} - \bar{y}_{nmp})^2\right\}$$

$$= E\left\{\sum^{n} (\bar{y}_{imp}^2 - \bar{y}_{nmp}^2)\right\}$$

$$= E\left\{\sum^{n} E(\bar{y}_{imp}^2 \mid i)\right\} - nE(\bar{y}_{nmp}^2)$$

Now, from (10), we obtain

$$E(\bar{y}_{imp}^2 \mid i) = \left\{\bar{y}_{i..}^2 + \left(\frac{1}{m} - \frac{1}{M}\right)\mathbf{S}_i^2 + \left(\frac{1}{p} - \frac{1}{P}\right)\frac{\bar{\mathbf{S}}_i^2}{m}\right\}$$

whence, taking further expectations and using (51), we get

$$\begin{aligned} E\{(n-1)\ s_b^2\} = \frac{n}{N}&\left\{\sum_{i=1}^{N} \bar{y}_{i..}^2 + \left(\frac{1}{m} - \frac{1}{M}\right)\sum_{i=1}^{N} \mathbf{S}_i^2\right. \\ &\left. + \left(\frac{1}{p} - \frac{1}{P}\right)\frac{1}{m}\sum_{i=1}^{N} \bar{\mathbf{S}}_i^2\right\} \\ &- n\left\{\left(\frac{1}{n} - \frac{1}{N}\right)\mathbf{S}_b^2 + \left(\frac{1}{m} - \frac{1}{M}\right)\frac{\bar{\mathbf{S}}_w^2}{n}\right. \\ &\left. + \left(\frac{1}{p} - \frac{1}{P}\right)\frac{\bar{\bar{\mathbf{S}}}_p^2}{nm} + \bar{y}_{...}^2\right\} \\ = (n-1)&\left\{\mathbf{S}_b^2 + \left(\frac{1}{m} - \frac{1}{M}\right)\bar{\mathbf{S}}_w^2\right. \\ &\left. + \left(\frac{1}{p} - \frac{1}{P}\right)\frac{\bar{\bar{\mathbf{S}}}_p^2}{m}\right\} \end{aligned}$$

or

$$E(s_b^2) = \mathbf{S}_b^2 + \left(\frac{1}{m} - \frac{1}{M}\right)\bar{\mathbf{S}}_v^2 + \left(\frac{1}{p} - \frac{1}{P}\right)\frac{\bar{\bar{\mathbf{S}}}_p^2}{m} \tag{55}$$

Substituting from (53) and (54) in (55), and from (53) in (54), we have

$$\text{Est. } \bar{\mathbf{S}}_w^2 = \bar{s}_w^2 - \left(\frac{1}{p} - \frac{1}{P}\right)\bar{\bar{s}}_p^2 \tag{56}$$

and

$$\text{Est. } \mathbf{S}_b^2 = s_b^2 - \left(\frac{1}{m} - \frac{1}{M}\right)\bar{s}_w^2 - \left(\frac{1}{p} - \frac{1}{P}\right)\frac{\bar{\bar{s}}_p^2}{M} \tag{57}$$

whence

$$\text{Est. } V(\bar{y}_{nmp}) = \left(\frac{1}{n} - \frac{1}{N}\right)s_b^2 + \left(\frac{1}{m} - \frac{1}{M}\right)\frac{\bar{s}_w^2}{N} + \left(\frac{1}{p} - \frac{1}{P}\right)\frac{\bar{\bar{s}}_p^2}{NM} \tag{58}$$

When N is large, the estimates of $\bar{\mathbf{S}}_w^2$ and $\mathbf{S}_b^2$ are still given by (56) and (57) but the estimate of the variance of the sample mean is given by the simple expression

$$\text{Est. } V(\bar{y}_{nmp}) = \frac{s_b^2}{n} \tag{59}$$

$$= \frac{\text{Mean square between first-stage units in the analysis of variance table for the sample}}{nmp}$$

When the other finite multipliers are also ignored, equations (54) to (57) reduce to

$$E(\bar{s}_w^2) = \bar{\mathbf{S}}_w^2 + \frac{1}{p}\bar{\bar{\mathbf{S}}}_p^2 \tag{60}$$

$$E(s_b^2) = \mathbf{S}_b^2 + \frac{\bar{\mathbf{S}}_w^2}{m} + \frac{\bar{\bar{\mathbf{S}}}_p^2}{mp} \tag{61}$$

$$\text{Est. } \bar{\mathbf{S}}_w^2 = \bar{s}_w^2 - \frac{1}{p}\bar{\bar{s}}_p^2 \tag{62}$$

$$\text{Est. } \mathbf{S}_b^2 = s_b^2 - \frac{\bar{s}_w^2}{m} \tag{63}$$

while the estimate of the variance of the sample mean is given by (59) as before.

7.10 Distribution of the Sample among the Three Stages

Lastly, we shall consider the problem of optimum allocation of the sample among the three stages when the cost of the survey is represented by

$$C = c_1 n + c_2 nm + c_3 nmp \tag{64}$$

where c_1, c_2 and c_3 are positive constants.

From (64) and (51), deleting the bars over $\mathbf{S}_w^2$ and $\mathbf{S}_p^2$ for convenience, we consider the product

$$C\left(V + \frac{\mathbf{S}_b^2}{N}\right)$$

$$= \left\{\left(\mathbf{S}_b^2 - \frac{1}{M}\mathbf{S}_w^2\right) + \left(\mathbf{S}_w^2 - \frac{1}{P}\mathbf{S}_p^2\right)\frac{1}{m} + \frac{1}{mp}\mathbf{S}_p^2\right\} \times \{c_1 + c_2 m + c_3 mp\}$$

which can also be written as

$$= \left\{c_1\left(\mathbf{S}_b^2 - \frac{1}{M}\mathbf{S}_w^2\right) + c_2\left(\mathbf{S}_w^2 - \frac{1}{P}\mathbf{S}_p^2\right) + c_3\mathbf{S}_p^2\right\}$$
$$+ \left\{c_2\left(\mathbf{S}_b^2 - \frac{1}{M}\mathbf{S}_w^2\right) m + \left(\mathbf{S}_w^2 - \frac{1}{P}\mathbf{S}_p^2\right)\frac{c_1}{m}\right\}$$
$$+ \left\{c_3\left(\mathbf{S}_w^2 - \frac{1}{P}\mathbf{S}_p^2\right) p + \frac{c_2\mathbf{S}_p^2}{p}\right\}$$
$$+ \left\{c_3\left(\mathbf{S}_b^2 - \frac{1}{M}\mathbf{S}_w^2\right) mp + \frac{c_1\mathbf{S}_p^2}{mp}\right\} \qquad (65)$$

The values of m and p which minimize (65) give the optimum allocation for both cases, when C is minimized for a given V_0, or when V is minimized for a given C_0.

When $(\mathbf{S}_b^2 - \mathbf{S}_w^2/M)$ and $(\mathbf{S}_w^2 - \mathbf{S}_p^2/P)$ are both positive, (65) can be rewritten as the sum of four square terms:

$$C\left(V + \frac{\mathbf{S}_b^2}{N}\right)$$

$$= \left\{\sqrt{c_1\left(\mathbf{S}_b^2 - \frac{1}{M}\mathbf{S}_w^2\right)} + \sqrt{c_2\left(\mathbf{S}_w^2 - \frac{1}{P}\mathbf{S}_p^2\right)} + \sqrt{c_3\mathbf{S}_p^2}\right\}^2$$
$$+ \left\{\sqrt{c_2\left(\mathbf{S}_b^2 - \frac{1}{M}\mathbf{S}_w^2\right)m} - \sqrt{\frac{c_1\left(\mathbf{S}_w^2 - \frac{1}{P}\mathbf{S}_p^2\right)}{m}}\right\}^2$$
$$+ \left\{\sqrt{c_3\left(\mathbf{S}_w^2 - \frac{1}{P}\mathbf{S}_p^2\right)p} - \sqrt{\frac{c_2\mathbf{S}_p^2}{p}}\right\}^2$$
$$+ \left\{\sqrt{c_3\left(\mathbf{S}_b^2 - \frac{1}{M}\mathbf{S}_w^2\right)mp} - \sqrt{\frac{c_1\mathbf{S}_p^2}{mp}}\right\}^2 \qquad (66)$$

Clearly, (66) is minimum when the last three square terms are all zero; then $\hat{m}$ is the nearest integer to

$$\sqrt{\frac{c_1\left(S_w^2 - \frac{1}{P}\, S_p^2\right)}{c_2\left(S_b^2 - \frac{1}{M} S_w^2\right)}} \tag{67}$$

and $\hat{p}$ is the nearest integer to

$$\sqrt{\frac{c_2\, S_p^2}{c_3\left(S_w^2 - \frac{1}{P}\, S_p\right)}} \tag{68}$$

It is interesting to note that the solution for p is independent of c_1 and S_b^2. Indeed, $\hat{p}$ bears the same relationship to c_2, c_3, S_w^2 and S_p^2 as $\hat{m}$ in (38) bears with c_1, c_2, S_b^2 and S_w^2, as one would, in fact, expect.

The above solutions presuppose that $S_b^2 - S_w^2/M$ and $S_w^2 - S_p^2/P$ are both positive, which may not be so.

Case I

Suppose $S_b^2 - S_w^2/M \leqslant 0$, but $S_w^2 - S_p^2/P > 0$.

In this case, to minimize (65), m must assume the maximum attainable value, say $\hat{m}$, which can also be equal to M, and p is to be such that

$$pc_3\left(S_w^2 - \frac{1}{P}\, S_p^2\right) + \frac{1}{p}\, c_2 S_p^2 + \hat{m} p c_3\left(S_b^2 - \frac{1}{M}\, S_w^2\right) + \frac{1}{p\hat{m}}\, c_1 S_p^2 \tag{69}$$

is minimum.

If

$$\left\{c_3\left(S_w^2 - \frac{1}{P}\, S_p^2\right) + c_3 \hat{m}\left(S_b^2 - \frac{1}{M}\, S_w^2\right)\right\} \leqslant 0$$

the expression (69) is minimum when p has the maximum attainable value.

In case this expression is positive, (69) is minimum when $\hat{p}$ is the nearest integer to

$$\sqrt{\frac{\left(\frac{1}{\hat{m}}\, c_1 + c_2\right) \mathrm{S}_p^2}{c_3\left(\mathrm{S}_w^2 - \frac{1}{P}\, \mathrm{S}_p^2\right) + \hat{m} c_3\left(\mathrm{S}_b^2 - \frac{1}{M}\, \mathrm{S}_w^2\right)}} \tag{70}$$

Case II

Suppose next that both $\mathrm{S}_b^2 - \mathrm{S}_w^2/M$ and $\mathrm{S}_w^2 - \mathrm{S}_p^2/P$ are negative or zero.

In this case (65) is minimum when p has the maximum attainable value, say $\hat{p}$, which could also be equal to P, and m is to be chosen so that

$$c_2\left(\mathrm{S}_b^2 - \frac{1}{M}\, \mathrm{S}_w^2\right) m + \left(\mathrm{S}_w^2 - \frac{1}{P}\, \mathrm{S}_p^2\right) \frac{c_1}{m} + m\hat{p} c_3\left(\mathrm{S}_b^2 - \frac{1}{M}\, \mathrm{S}_w^2\right) + \frac{1}{\hat{p} m}\, c_1 \mathrm{S}_p^2 \tag{71}$$

is minimum. Now (71) can be rewritten as

$$m\left(\mathrm{S}_b^2 - \frac{1}{M}\, \mathrm{S}_w^2\right)(c_2 + \hat{p} c_3) + \frac{c_1}{m}\left(\frac{1}{\hat{p}}\, \mathrm{S}_p^2 - \frac{1}{P}\, \mathrm{S}_p^2 + \mathrm{S}_w^2\right) \tag{72}$$

which is minimum when m has the maximum attainable value.

Case III

Suppose now that $\mathrm{S}_b^2 - \mathrm{S}_w^2/M > 0$, but $\mathrm{S}_w^2 - \mathrm{S}_p^2/P \leqslant 0$.

The solution in this case for pm is given by the nearest integer to

$$\sqrt{\frac{c_1\, \mathrm{S}_p^2}{c_3\left(\mathrm{S}_b^2 - \frac{1}{M}\, \mathrm{S}_w^2\right)}} \tag{73}$$

with p attaining the maximum and m the minimum possible values.

7.11 Two-Stage Sampling, Unequal First-Stage Units: Estimate of the Population Mean

We shall now give the theory appropriate for unequal first-stage units. We shall assume selection with equal probability at each stage of selection.

Let

M_i = the number of second-stage units in the i-th first-stage unit $(i = 1, 2, \ldots, N)$

m_i = the number of second-stage units to be selected from the i-th first-stage unit, if in the sample

M_0 = the total number of second-stage units in the population

$$i.e., \quad \sum_{i=1}^{N} M_i$$

m_0 = the number of second-stage units in the sample

$$i.e., \quad \sum^{n} m_i \quad \text{or} \quad \sum_{i=1}^{N} a_i m_i$$

where $a_i = 1$, if the i-th first-stage unit is in the sample, and otherwise zero.

$$\bar{y}_{i(m_i)} = \frac{1}{m_i} \sum_{j}^{m_i} y_{ij}$$

$$\bar{y}_{n(m_i)} = \frac{1}{n} \sum^{n} \bar{y}_{i(m_i)} = \frac{1}{n} \sum^{n} \frac{1}{m_i} \sum_{j}^{m_i} y_{ij}$$

and

$$\bar{y}_{m_0} = \frac{1}{m_0} \sum_{i}^{n} \sum_{j}^{m_i} y_{ij}$$

Further, let

$$\bar{y}_{i.} = \bar{y}_{i(M_i)} = \frac{1}{M_i} \sum_{j=1}^{M_i} y_{ij}$$

$$\bar{y}_{N.} = \bar{y}_{N(M_i)} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{M_i} \sum_{j=1}^{M_i} y_{ij}$$

and

$$\bar{y}_{..} = \bar{y}_{M_0} = \frac{1}{M_0}\sum_{i=1}^{N}\sum_{j=1}^{M_i} y_{ij}$$

Now several estimates of the population mean $\bar{y}_{..}$ can be formed. The simplest is the simple mean of the first-stage unit means which we shall denote as $\bar{y}_s$, given by

$$\bar{y}_s = \bar{y}_{n(m_i)}$$

$$= \frac{1}{n}\sum^{n} \bar{y}_{i(m_i)}$$

where the summation extends over the units in the sample. This can also be written as

$$\bar{y}_s = \frac{1}{n}\sum_{i=1}^{N} a_i\,\bar{y}_{i(m_i)} \tag{74}$$

where a_i is a random variable such that $a_i = 1$ if the i-th first-stage unit is in the sample, and otherwise zero.

A second estimate which we shall denote as $\bar{y}_s'$ is based on the first-stage unit totals, given by

$$\bar{y}_s' = \bar{y}'_{n(m_i)}$$

$$= \frac{1}{n\bar{M}}\sum^{n} M_i\,\bar{y}_{i(m_i)}$$

and can be written as

$$\bar{y}_s' = \frac{1}{n\bar{M}}\sum_{i=1}^{N} a_i\,(M_i\,\bar{y}_{i(m_i)}) \tag{75}$$

where

$M_i\,\bar{y}_{i(m_i)}$ is the estimate of the population total for the i-th first-stage unit

and

$a_i = 1$ if the i-th cluster is in the sample, and zero otherwise.

Yet another estimate which we shall denote as $\bar{y}_s''$ is the ratio estimate, defined by

$$\bar{y}_s'' = \frac{\sum^n M_i \bar{y}_{i(m_i)}}{\sum^n M_i}$$

and can be written as simply

$$\bar{y}_s'' = \frac{\bar{y}_s'}{\bar{u}_n} \tag{76}$$

where

$$u_i = \frac{M_i}{\bar{M}} \quad \text{and} \quad \bar{u}_n = \frac{1}{n}\sum^n u_i \quad .$$

More generally, we shall form a ratio estimate of the population mean.

Let

x be a supplementary variate

$\bar{x}_{..}$ be the population mean, assumed known

R be the population ratio $\frac{\bar{y}_{..}}{\bar{x}_{..}}$

$$\bar{x}_s' = \frac{1}{n\bar{M}}\sum^n M_i \bar{x}_{i(m_i)}$$

and

$$R_s = \frac{\bar{y}_s'}{\bar{x}_s'}$$

Then the ratio estimate of the population mean $\bar{y}_{..}$ is defined by

$$\bar{y}_R = R_s \bar{x}_{..} \tag{77}$$

We shall study the properties of the different estimates in the next section.

7.12 Two-Stage Sampling, Unequal First-Stage Units: Expected Values and Variances of the Different Estimates

(*a*) *Estimate* $\bar{y}_s$

We write, from (74),

$$E(\bar{y}_s) = E\left\{\frac{1}{n}\sum_{i=1}^{N} a_i\, E(\bar{y}_{i(m_i)} \mid i)\right\}$$

$$= E\left\{\frac{1}{n}\sum_{i=1}^{N} a_i\, \bar{y}_{i.}\right\}$$

$$= \frac{1}{n}\sum_{i=1}^{N} E(a_i)\cdot \bar{y}_{i.} \qquad (78)$$

Now, by definition, $a_i = 1$ if the *i*-th unit is in the sample, and is otherwise zero. Clearly, $E(a_i)$ is the probability of including the *i*-th first-stage unit in the sample. We have seen in Chapter II that, when units are selected with equal probability, $E(a_i)$ is n/N. We therefore have from (78),

$$E(\bar{y}_s) = \frac{1}{N}\sum_{i=1}^{N} \bar{y}_{i.}$$

$$= \bar{y}_{N.}$$

$$\neq \bar{y}_{..} \qquad (79)$$

thus showing that $\bar{y}_s$ is a biased estimate of the population mean. We note that the probability of including a specified unit in the sample is independent of the unit when the selection probabilities are equal. In evaluating the expected values it is not, therefore, necessary to introduce the random variable a, the use of the theorem of Section 2*a*.9 on expectations in Chapter II being sufficient for the purpose. Thus,

$$E(\bar{y}_s) = E\left\{\frac{1}{n}\sum^{n} E(\bar{y}_{i(m_i)} \mid i)\right\}$$

$$= E\left\{\frac{1}{n}\sum^{n} \bar{y}_{i.}\right\}$$

$$= \frac{1}{N}\sum_{i=1}^{N} \bar{y}_{i.}$$

$$= \bar{y}_{N.}$$

To evaluate the mean square error of $\bar{y}_S$, we write

$$\bar{y}_s - \bar{y}_{..} = \bar{y}_s - \bar{y}_{n.} + \bar{y}_{n.} - \bar{y}_{N.} + \bar{y}_{N.} - \bar{y}_{..}$$

whence, squaring both sides and taking expectations term by term, we get

$$\begin{aligned} M.S.E.\,(\bar{y}_s) = {} & E\,(\bar{y}_s - \bar{y}_{n.})^2 + E\,(\bar{y}_{n.} - \bar{y}_{N.})^2 + E\,(\bar{y}_{N.} - \bar{y}_{..})^2 \\ & + 2E\,(\bar{y}_s - \bar{y}_{n.})\,(\bar{y}_{n.} - \bar{y}_{N.}) \\ & + 2E\,(\bar{y}_s - \bar{y}_{n.})\,(\bar{y}_{N.} - \bar{y}_{..}) \\ & + 2E\,(\bar{y}_{n.} - \bar{y}_{N.})\,(\bar{y}_{N.} - \bar{y}_{..}) \end{aligned} \tag{80}$$

Taking the first term in (80), we have

$$\begin{aligned} E\,(\bar{y}_s - \bar{y}_{n.})^2 &= E\left\{\frac{1}{n}\sum^{n} (\bar{y}_{i(m_i)} - \bar{y}_{i.})\right\}^2 \\ &= \frac{1}{n^2}\,E\left[\sum^{n} E\{(\bar{y}_{i(m_i)} - \bar{y}_{i.})^2 \mid i\}\right. \\ &\qquad \left. + \sum_{i \neq i'}^{n} E\{(\bar{y}_{i(m_i)} - \bar{y}_{i.})\,(\bar{y}_{i'(m_{i'})} - \bar{y}_{i'.}) \mid i, i'\}\right] \\ &= \frac{1}{n^2}\,E\left\{\sum^{n}\left(\frac{1}{m_i} - \frac{1}{M_i}\right) S_i^2 + 0\right\} \\ &= \frac{1}{nN}\sum_{i=1}^{N}\left(\frac{1}{m_i} - \frac{1}{M_i}\right) S_i^2 \end{aligned} \tag{81}$$

where

$$\mathrm{S}_i^2 = \frac{\sum\limits_{j=1}^{M_i} (y_{ij} - \bar{y}_{i.})^2}{M_i - 1}$$

The value of the second term in (80) is obviously given by

$$E\,(\bar{y}_{n.} - \bar{y}_{N.})^2 = \left(\frac{1}{n} - \frac{1}{N}\right) \mathrm{S}_b^2 \tag{82}$$

where

$$\mathrm{S}_b^2 = \frac{1}{N-1} \sum_{i=1}^{N} (\bar{y}_{i.} - \bar{y}_{N.})^2$$

The third term in (80) is unaffected by the sampling procedure and therefore a constant, being the square of the bias term in the estimate. The fourth term is clearly zero since the procedure of selecting a sub-sample from a given first-stage unit is independent of the procedure of drawing a sample of first-stage units. We have

$$\begin{aligned} E\,(\bar{y}_s - \bar{y}_{n,})\,(\bar{y}_{n.} - \bar{y}_{N.}) &= E\left\{\frac{1}{n} \sum^{n} E\,\{(\bar{y}_{i(m_i)} - \bar{y}_{i.}) \mid i\}\,(\bar{y}_{n.} - \bar{y}_{N.})\right\} \\ &= 0 \end{aligned} \tag{83}$$

The fifth and the sixth terms are obviously zero. We are therefore left with

$$\begin{aligned} M.S.E.\,(\bar{y}_s) &= \left(\frac{1}{n} - \frac{1}{N}\right) \mathrm{S}_b^2 + \frac{1}{nN} \sum_{i=1}^{N} \left(\frac{1}{m_i} - \frac{1}{M_i}\right) \mathrm{S}_i^2 \\ &\qquad\qquad + (\bar{y}_{N.} - \bar{y}_{..})^2 \\ &= V\,(\bar{y}_s) + (\bar{y}_{N.} - \bar{y}_{..})^2 \end{aligned} \tag{84}$$

where

$$\begin{aligned} V\,(\bar{y}_s) &= E\,(\bar{y}_s - \bar{y}_{N.})^2 \\ &= \left(\frac{1}{n} - \frac{1}{N}\right) \mathrm{S}_b^2 + \frac{1}{nN} \sum_{i=1}^{N} \left(\frac{1}{m_i} - \frac{1}{M_i}\right) \mathrm{S}_i^2 \end{aligned} \tag{85}$$

The bias component arises because of the failure to give equal chance to each second-stage unit being included in the sample. Unless the M_i vary considerably and the character is correlated with M_i, this component may not be serious; but it is important to collect evidence on this point before adopting $\bar{y}_s$ as the estimate of $\bar{y}_{..}$. The procedure of estimation in the yield surveys in India provides an instance in point (Sukhatme and Panse, 1951). The number of fields under the crop is often found to vary considerably from one village to another, thereby pointing to the need for testing the nature and magnitude of the bias arising from the use of the simple arithmetic mean of plot yields to estimate the average yield per plot. Table 7.4 gives a comparison of the yield estimates in large samples based on $\bar{y}_s$ with those calculated from an alternative estimate $\bar{y}_s'$ which, as will be presently shown, provides an unbiased estimate of the yield per unit. The standard error of the difference between the two estimates is known to be

TABLE 7.4

Mean Yield in the Wheat Survey in Punjab, 1943–44

District		Yield in Lb. per Acre	
		Simple Arithmetic Mean $\bar{y}_s$	Weighted Mean $\bar{y}_s'$
1. Amritsar	..	1029	1041
2. Gurdaspur	..	829	862
3. Jullundur	..	839	881
4. Hoshiarpur	..	804	796
5. Ludhiana	..	1247	1246
6. Ferozepur	..	1052	1079
7. Ambala	..	854	820
8. Karnal	..	839	868
9. Hissar	..	1090	1142
10. Rohtak	..	1004	997
11. Gurgaon	..	766	752
Province	..	920	927

of the order of 6 to 8%. It will be seen that not only do the differences change sign from district to district but also their magnitude is negligible compared to their standard errors.

(*b*) *Estimate* $\bar{y}_s'$

The estimate corresponds to $\bar{y}_{n.}'$ of the previous chapter and like $\bar{y}_{n.}'$ provides an unbiased estimate of the population mean. For,

$$E(\bar{y}_s') = E\left\{\frac{1}{n\bar{M}}\sum^{n} E(M_i\bar{y}_{i(m_i)} \mid i)\right\}$$

$$= E\left\{\frac{1}{n\bar{M}}\sum^{n} M_i\bar{y}_{i.}\right\}$$

$$= \frac{1}{N\bar{M}}\sum_{i=1}^{N} M_i\bar{y}_{i.}$$

$$= \bar{y}_{..} \tag{86}$$

To obtain the sampling variance of $\bar{y}_s'$, we write

$$V(\bar{y}_s') = E(\bar{y}_s' - \bar{y}_{..})^2$$

$$= E(\bar{y}_s' - \bar{y}_{n.}' + \bar{y}_{n.}' - \bar{y}_{..})^2$$

$$= E(\bar{y}_s' - \bar{y}_{n.}')^2 + E(\bar{y}_{n.}' - \bar{y}_{..})^2 + 2E(\bar{y}_s' - \bar{y}_{n.}')(\bar{y}_{n.}' - \bar{y}_{..}) \tag{87}$$

Taking the first term in (87), we have

$$E(\bar{y}_s' - \bar{y}_{n.}')^2 = E\left\{\frac{1}{n\bar{M}}\sum^{n} M_i(\bar{y}_{i(m_i)} - \bar{y}_{i.})\right\}^2$$

$$= \frac{1}{n^2\bar{M}^2} E\left[\sum^{n} M_i^2 E\{(\bar{y}_{i(m_i)} - \bar{y}_{i.})^2 \mid i\} + \sum_{i \neq i'}^{n} M_i M_{i'} E\{(\bar{y}_{i(m_i)} - \bar{y}_{i.})(\bar{y}_{i'(m_{i'})} - \bar{y}_{i'.}) \mid i, i'\}\right]$$

$$= \frac{1}{n^2\bar{M}^2}\, E\left[\sum^{n} M_i^2 \left(\frac{1}{m_i} - \frac{1}{M_i}\right) \mathbf{S}_i^2\right]$$

$$= \frac{1}{nN\bar{M}^2} \sum_{i=1}^{N} M_i^2 \left(\frac{1}{m_i} - \frac{1}{M_i}\right) \mathbf{S}_i^2 \tag{88}$$

Also, we know that

$$E(\bar{y}_{n.}' - \bar{y}_{..})^2 = \left(\frac{1}{n} - \frac{1}{N}\right) \mathbf{S}_b'^2 \tag{89}$$

where

$$\mathbf{S}_b'^2 = \frac{1}{N-1} \sum_{i=1}^{N} \left(\frac{M_i}{\bar{M}}\,\bar{y}_{i.} - \bar{y}_{..}\right)^2$$

$$= \frac{1}{N-1} \sum_{i=1}^{N} (u_i\bar{y}_{i.} - \bar{y}_{..})^2 \tag{90}$$

The last term in (87) is clearly zero and we are left with

$$V(\bar{y}_s') = \left(\frac{1}{n} - \frac{1}{N}\right) \mathbf{S}_b'^2 + \frac{1}{nN} \sum_{i=1}^{N} \frac{M_i^2}{\bar{M}^2} \left(\frac{1}{m_i} - \frac{1}{M_i}\right) \mathbf{S}_i^2 \tag{91}$$

It will be noticed that the first component depends upon the variation between the cluster totals. It can be shown, in general, to be larger than the corresponding component in (85), provided the correlation between the cluster size and the cluster mean is positive and the bias component $(\bar{y}_{N.} - \bar{y}_{..})$ is negligible. The second component of (91) is also likely to be larger than the corresponding component of (85) as there is likely to be positive correlation between M_i and $\mathbf{S}_i^2$. Unless the bias in $\bar{y}_s$ is therefore likely to be serious, the estimate $\bar{y}_s'$ may not be preferred to $\bar{y}_s$.

(c) *Ratio Estimate $\bar{y}_s''$*

We shall assume that the number of first-stage units in the sample is large enough to neglect the bias term in the expected value of a ratio estimate. To a first approximation, then

$$E(\bar{y}_s'') = \frac{E(\bar{y}_s')}{E(\bar{u}_n)}$$

$$= \bar{y}_{..} \tag{92}$$

since

$$E(\bar{u}_n) = 1$$

To derive the sampling variance of $\bar{y}_S''$, we make use of the result (28) of Chapter IV and write, since $\bar{u}_N = 1$, to a first approximation

$$V(\bar{y}_s'') = V(\bar{y}_s') + \bar{y}_{..}^2\, V(\bar{u}_n) - 2\bar{y}_{..}\, \text{Cov}\,(\bar{y}_s', \bar{u}_n) \tag{93}$$

Now, $V(\bar{y}_S')$ is known from (91); $V(\bar{u}_n)$ is given by

$$V(\bar{u}_n) = \left(\frac{1}{n} - \frac{1}{N}\right) \mathbf{S}_u^2 \tag{94}$$

where

$$\mathbf{S}_u^2 = \frac{1}{N-1} \sum_{i=1}^{N} (u_i - 1)^2$$

and

$$\text{Cov}\,(\bar{y}_s', \bar{u}_n) = E\,\{(\bar{u}_n - 1)\,(\bar{y}_s' - \bar{y}_{..})\}$$

$$= E\,[(\bar{u}_n - 1)\, E\,\{(\bar{y}_s' - \bar{y}_{n.}') \mid i\}]$$

$$+ E\,\{(\bar{y}_{n.}' - \bar{y}_{..})\,(\bar{u}_n - 1)\}$$

$$= \left(\frac{1}{n} - \frac{1}{N}\right) \frac{1}{N-1} \left\{\sum_{i=1}^{N} (u_i\bar{y}_{i.} - \bar{y}_{..})\,(u_i - 1)\right\} \tag{95}$$

Collecting together the terms, we get

$$V(\bar{y}_s'') = \left(\frac{1}{n} - \frac{1}{N}\right) \frac{1}{N-1} \left\{\sum_{i=1}^{N} (u_i\bar{y}_{i.} - \bar{y}_{..})^2 + \bar{y}_{..}^2 \sum_{i=1}^{N} (u_i - 1)^2 \right.$$

$$\left. - 2\bar{y}_{..} \sum_{i=1}^{N} (u_i\bar{y}_{i.} - \bar{y}_{..})\,(u_i - 1)\right\}$$

$$+ \frac{1}{nN} \sum_{i=1}^{N} u_i^2 \left(\frac{1}{m_i} - \frac{1}{M_i}\right) \mathbf{S}_i^2$$

which, on simplification, gives

$$V(\bar{y}_s'') = \left(\frac{1}{n} - \frac{1}{N}\right) \mathbf{S}_b''^2 + \frac{1}{nN} \sum_{i=1}^{N} u_i^2 \left(\frac{1}{m_i} - \frac{1}{M_i}\right) \mathbf{S}_i^2 \qquad (96)$$

where

$$\mathbf{S}_b''^2 = \frac{1}{N-1} \sum_{i=1}^{N} u_i^2 (\bar{y}_{i.} - \bar{y}_{..})^2$$

It will be noticed that the second term of (96) is identical with the second term of (91); the first term, on the other hand, is expected to be less than the corresponding term of (91), if $M_i\bar{y}_{i.}$ and M_i are positively correlated and the correlation is greater than one-half. It will, however, in general, be larger than the corresponding term in (85), provided M_i and $(\bar{y}_{i.} - \bar{y}_{..})$ are positively correlated and the bias $(\bar{y}_{N.} - \bar{y}_{..})$ is negligible. Data on the estimation of yield of crops provide an example of the relative efficiency of the three estimates $\bar{y}_S$, $\bar{y}_S'$ and $\bar{y}_S''$. Table 7.5 gives the percentage standard errors of the three estimates, made on four different yield surveys: wheat survey in U.P. during 1947–48; wheat survey in Delhi during 1948–49; and cotton surveys in Madhya Pradesh during 1944–45 and 1945–46. It is seen that the unweighted mean of plot yields $(\bar{y}_S)$ has the least standard error, considerably less than those of the other two estimates. This estimate is, of course, biased; but the bias, as we saw in Table 7.4, which is typical of these surveys, is found to be negligible for all practical purposes. In crop surveys in India,

TABLE 7.5

Percentage Standard Errors of Different Estimates of Mean Yield

	$\bar{y}_s$	$\bar{y}_s'$	$\bar{y}_s''$
Wheat (U.P.), 1947–48	3·7	14·0	4·7
Wheat (Delhi), 1948–49	2·5	10·0	5·7
Cotton (Madhya Pradesh), 1944–45 ..	5·5	15·0	11·3
1945–46 ..	6·9	14·0	13·2

the M_i's are found to vary considerably from village to village, with the result that the estimate $\bar{y}_S'$ turns out to be markedly inefficient, as shown in Table 7.5.

(*d*) *Ratio Estimate* $\bar{y}_R$

We shall assume that n is large enough to ignore the bias terms of the first and higher orders in the expected value of the estimate $\bar{y}_R$. To obtain the variance, we write using (28) of Chapter IV,

$$V_1(\bar{y}_R) = \bar{y}_{..}^2\left\{\frac{V(\bar{y}_s')}{\bar{y}_{..}^2} + \frac{V(\bar{x}_s')}{\bar{x}_{..}^2} - \frac{2\,\text{Cov}\,(\bar{y}_s', \bar{x}_s')}{\bar{y}_{..}\bar{x}_{..}}\right\} \tag{97}$$

Now, by analogy with (91),

$$V(\bar{x}_s') = \left(\frac{1}{n} - \frac{1}{N}\right) \text{S}_{bx}'^2 + \frac{1}{nN}\sum_{i=1}^{N} u_i^2\left(\frac{1}{m_i} - \frac{1}{M_i}\right)\text{S}_{ix}^2 \tag{98}$$

$V(\bar{y}_S')$ is known from (91) and may be written as

$$V(\bar{y}_s') = \left(\frac{1}{n} - \frac{1}{N}\right) \text{S}_{by}'^2 + \frac{1}{nN}\sum_{i=1}^{N} u_i^2\left(\frac{1}{m_i} - \frac{1}{M_i}\right)\text{S}_{iy}^2 \tag{99}$$

and

$$\begin{aligned}\text{Cov}\,(\bar{y}_s', \bar{x}_s') &= E\{(\bar{y}_s' - \bar{y}_{..})(\bar{x}_s' - \bar{x}_{..})\}\\ &= E\{(\bar{y}_s' - \bar{y}_{n.}' + \bar{y}_{n.}' - \bar{y}_{..})\\ &\qquad\qquad \times (\bar{x}_s' - \bar{x}_{n.}' + \bar{x}_{n.}' - \bar{x}_{..})\}\\ &= E\{(\bar{y}_s' - \bar{y}_{n.}')(\bar{x}_s' - \bar{x}_{n.}')\\ &\qquad\qquad + (\bar{y}_{n.}' - \bar{y}_{..})(\bar{x}_{n.}' - \bar{x}_{..})\}\end{aligned} \tag{100}$$

since expectations of the other two product terms are zero. Taking the first term in (100), we have

$$\begin{aligned}&E\{(\bar{y}_s' - \bar{y}_{n.}')(\bar{x}_s' - \bar{x}_{n.}')\}\\ &\qquad = E\left[\left\{\frac{1}{n}\sum^{n} u_i\,(\bar{y}_{i(m_i)} - \bar{y}_{i.})\right\}\left\{\frac{1}{n}\sum^{n} u_i\,(\bar{x}_{i(m_i)} - \bar{x}_{i.})\right\}\right]\end{aligned}$$

$$= \frac{1}{n^2} E\left[\sum^{n} u_i^2 (\bar{y}_{i(m_i)} - \bar{y}_{i.})(\bar{x}_{i(m_i)} - \bar{x}_{i.}) + \sum_{i \neq i'}^{n} u_i u_{i'} (\bar{y}_{i(m_i)} - \bar{y}_{i.})(\bar{x}_{i'(m_{i'})} - \bar{x}_{i'.})\right]$$

$$= \frac{1}{n^2} E\left[\sum^{n} u_i^2 E\{(\bar{y}_{i(m_i)} - \bar{y}_{i.})(\bar{x}_{i(m_i)} - \bar{x}_{i.}) \mid i\}\right] + \frac{1}{n^2} E\left[\sum_{i \neq i'}^{n} u_i u_{i'} E\{(\bar{y}_{i(m_i)} - \bar{y}_{i.})(\bar{x}_{i'(m_{i'})} - \bar{x}_{i'.}) \mid i, i'\}\right]$$

$$= \frac{1}{n^2} E\left[\sum^{n} u_i^2 \left(\frac{1}{m_i} - \frac{1}{M_i}\right) \mathbf{S}_{iyx}\right]$$

$$= \frac{1}{nN} \sum_{i=1}^{N} u_i^2 \left(\frac{1}{m_i} - \frac{1}{M_i}\right) \mathbf{S}_{iyx} \tag{101}$$

where

$$\mathbf{S}_{iyx} = \frac{1}{M_i - 1} \sum_{j=1}^{M_i} \{(y_{ij} - \bar{y}_{i.})(x_{ij} - \bar{x}_{i.})\} \tag{102}$$

The second term in (100) gives

$$E(\bar{y}_{n.}' - \bar{y}_{..})(\bar{x}_{n.}' - \bar{x}_{..}) = \left(\frac{1}{n} - \frac{1}{N}\right) \mathbf{S}_{byx}' \tag{103}$$

where

$$\mathbf{S}_{byx}' = \frac{1}{N-1} \sum_{i=1}^{N} (u_i \bar{y}_{i.} - \bar{y}_{..})(u_i \bar{x}_{i.} - \bar{x}_{..}) \tag{104}$$

so that from (100), (101) and (103),

$$\text{Cov}\,(\bar{y}_s', \bar{x}_s') = \left(\frac{1}{n} - \frac{1}{N}\right) \mathbf{S}_{byx}' + \frac{1}{nN} \sum_{i=1}^{N} u_i^2 \left(\frac{1}{m_i} - \frac{1}{M_i}\right) \mathbf{S}_{iyx} \tag{105}$$

Substituting in (97) from (98), (99) and (105), we get

$$
\begin{aligned}
V_1(\bar{y}_R) = {} & \left(\frac{1}{n} - \frac{1}{N}\right) \left\{ S_{by}'^2 + \frac{\bar{y}_{..}^2}{\bar{x}_{..}^2} S_{bx}'^2 - \frac{2\bar{y}_{..}}{\bar{x}_{..}} S_{byx}' \right\} \\
& + \frac{1}{nN} \left\{ \sum_{i=1}^{N} u_i^2 \left(\frac{1}{m_i} - \frac{1}{M_i}\right) S_{iy}^2 \right. \\
& + \frac{\bar{y}_{..}^2}{\bar{x}_{..}^2} \sum_{i=1}^{N} u_i^2 \left(\frac{1}{m_i} - \frac{1}{M_i}\right) S_{ix}^2 \\
& \left. - \frac{2\bar{y}_{..}}{\bar{x}_{..}} \sum_{i=1}^{N} u_i^2 \left(\frac{1}{m_i} - \frac{1}{M_i}\right) S_{iyx} \right\} \\
= {} & \left(\frac{1}{n} - \frac{1}{N}\right) \frac{1}{N-1} \left\{ \left(\sum_{i=1}^{N} u_i^2 \bar{y}_{i.}^2 - N\bar{y}_{..}^2 \right) \right. \\
& + \frac{\bar{y}_{..}^2}{\bar{x}_{..}^2} \left(\sum_{i=1}^{N} u_i^2 \bar{x}_{i.}^2 - N\bar{x}_{..}^2 \right) \\
& \left. - \frac{2\bar{y}_{..}}{\bar{x}_{.}} \left(\sum_{i=1}^{N} u_i^2 \bar{y}_{i.} \bar{x}_{i.} - N\bar{y}_{..}\bar{x}_{..} \right) \right\} \\
& + \frac{1}{nN} \left\{ \sum_{i=1}^{N} u_i^2 \left(\frac{1}{m_i} - \frac{1}{M_i}\right) \left(S_{iy}^2 + \frac{\bar{y}_{..}^2}{\bar{x}_{..}^2} S_{ix}^2 \right. \right. \\
& \left. \left. - \frac{2\bar{y}_{..}}{\bar{x}_{..}} S_{iyx} \right) \right\} \\
= {} & \left(\frac{1}{n} - \frac{1}{N}\right) \frac{1}{N-1} \sum_{i=1}^{N} u_i^2 (\bar{y}_{i.} - R\bar{x}_{i.})^2 \\
& + \frac{1}{nN} \sum_{i=1}^{N} u_i^2 \left(\frac{1}{m_i} - \frac{1}{M_i}\right) D_i^2 \qquad (106)
\end{aligned}
$$

where

$$D_i^2 = \mathbf{S}_{iy}^2 + R^2\mathbf{S}_{ix}^2 - 2R\mathbf{S}_{iyx} \tag{107}$$

When $x_{ij} = 1$, (107) becomes

$$D_i^2 = \mathbf{S}_{iy}^2$$

and the variance (106) reduces to the expression (96) as expected.

In general, when M_i's vary rather considerably the estimate $\bar{y}_R$ is likely to be the most efficient of the four estimates, provided n is large and x is highly correlated with y.

7.13 Two-Stage Sampling, Unequal First-Stage Units: Estimation of the Variances from the Sample

(a) Mean of Cluster Means Estimate $\bar{y}_s$

Consider the mean square between cluster means in the sample, s_b^2 as defined by

$$s_b^2 = \frac{\sum^n (\bar{y}_{i(mi)} - \bar{y}_s)^2}{n-1}$$

$$= \frac{\sum^n \bar{y}^2_{i(mi)} - n\bar{y}_s^2}{n-1} \tag{108}$$

Multiplying both sides of (108) by $(n - 1)$ and taking expectations, we obtain

$$(n-1)\,E\,(s_b^2) = E\left\{\sum^n E\,(\bar{y}^2_{i(mi)} \mid i)\right\} - nE\,(\bar{y}_s^2)$$

$$= E\left[\sum^n \left\{\bar{y}_{i.}^2 + \left(\frac{1}{m_i} - \frac{1}{M_i}\right)\mathbf{S}_i^2\right\}\right] - nE\,(\bar{y}_s^2)$$

$$= \frac{n}{N}\sum_{i=1}^{N} \left\{\bar{y}_{i.}^2 + \left(\frac{1}{m_i} - \frac{1}{M_i}\right)\mathbf{S}_i^2\right\}$$

$$- n\,\{\bar{y}_{N.}^2 + V\,(\bar{y}_s)\}$$

whence, substituting from (85) and remembering that

$$\mathrm{S}_b^2 = \frac{1}{N-1} \sum_{i=1}^{N} (\bar{y}_{i.} - \bar{y}_{N.})^2$$

we obtain

$$E(s_b^2) = \mathrm{S}_b^2 + \frac{1}{N} \sum_{i=1}^{N} \left(\frac{1}{m_i} - \frac{1}{M_i}\right) \mathrm{S}_i^2 \tag{109}$$

Also

$$E\left\{\frac{1}{n} \sum^{n} \left(\frac{1}{m_i} - \frac{1}{M_i}\right) s_i^2\right\} = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{1}{m_i} - \frac{1}{M_i}\right) \mathrm{S}_i^2 \tag{110}$$

where

$$s_i^2 = \frac{1}{m_i - 1} \sum_{j}^{m_i} (y_{ij} - \bar{y}_{i(m_i)})^2 \tag{111}$$

Hence

$$\text{Est. } \mathrm{S}_b^2 = s_b^2 - \frac{1}{n} \sum^{n} \left(\frac{1}{m_i} - \frac{1}{M_i}\right) s_i^2 \tag{112}$$

Substituting from (110) and (112) in (85), we obtain

$$\text{Est. } V(\bar{y}_s) = \left(\frac{1}{n} - \frac{1}{N}\right) s_b^2 + \frac{1}{nN} \sum^{n} \left(\frac{1}{m_i} - \frac{1}{M_i}\right) s_i^2 \tag{113}$$

It should be pointed out that s_b^2 as defined in (108) cannot be derived from the mean square between clusters in the table showing the analysis of variance on y_{ij} unless m_i's are all equal. In most surveys, however, as will be shown in the next section, m_i will have a constant value within a stratum, although varying from stratum to stratum. Consequently, if analysis of variance is carried out separately for each stratum, we can estimate the variance of the mean for that stratum by substituting from the analysis of variance table. Thus, if B is the mean square between clusters and W the mean square within clusters in the sample from any stratum, and further, the within-variance S_i^2 is assumed to be constant for all i, we obtain

$$\text{Est. } V(\bar{y}_s) = \left(\frac{1}{n} - \frac{1}{N}\right) \frac{B}{m} + \frac{1}{N} \left(\frac{1}{m} - \frac{1}{\bar{M}_h}\right) W$$

where $\bar{M}_h$ is the harmonic mean of the M_i's in the sample. When $M_i = M$ ($i = 1, 2, \ldots, N$), and $m_i = m$ ($i = 1, 2, \ldots, N$), the formulæ become identical with those in Section 7.4.

For N large, the variance can be computed from the simple expression

$$\text{Est. } V(\bar{y}_s) = \frac{s_b^2}{n} \tag{114}$$

(*b*) *Mean of Cluster Totals Estimate* $\bar{y}_s'$

Consider the mean square per element basis of the cluster totals in the sample defined by

$$s_b'^2 = \frac{1}{n-1} \sum^{n} \left(\frac{M_i}{\bar{M}} \bar{y}_{i(m_i)} - \bar{y}_s' \right)^2 \tag{115}$$

On expanding and taking expectations, we get

$$(n-1)\, E(s_b'^2) = E \left\{ \sum^{n} \frac{M_i^2}{\bar{M}^2} E(\bar{y}^2{}_{i(m_i)} \mid i) \right\} - nE(\bar{y}_s'^2)$$

$$= \frac{n}{N} \sum_{i=1}^{N} \frac{M_i^2}{\bar{M}^2} \left\{ \bar{y}_{i.}^2 + \left(\frac{1}{m_i} - \frac{1}{M_i} \right) \mathbf{S}_i^2 \right\}$$
$$- n \{ \bar{y}_{..}^2 + V(\bar{y}_s') \}$$

Substituting from (91), we obtain

$$E(s_b'^2) = \mathbf{S}_b'^2 + \frac{1}{N} \sum_{i=1}^{N} \frac{M_i^2}{\bar{M}^2} \left(\frac{1}{m_i} - \frac{1}{M_i} \right) \mathbf{S}_i^2 \tag{116}$$

Also

$$E \left\{ \frac{1}{n} \sum^{n} \frac{M_i^2}{\bar{M}^2} \left(\frac{1}{m_i} - \frac{1}{M_i} \right) s_i^2 \right\} = \frac{1}{N} \sum_{i=1}^{N} \frac{M_i^2}{\bar{M}^2} \left(\frac{1}{m_i} - \frac{1}{M_i} \right) \mathbf{S}_i^2 \tag{117}$$

Hence

$$\text{Est. } \mathbf{S}_b'^2 = s_b'^2 - \frac{1}{n} \sum^{n} \frac{M_i^2}{\bar{M}^2} \left(\frac{1}{m_i} - \frac{1}{M_i} \right) s_i^2 \tag{118}$$

On substituting in (91) from (117) and (118), we get

$$\text{Est. } V(\bar{y}_s') = \left(\frac{1}{n} - \frac{1}{N}\right) s_b'^2 + \frac{1}{nN} \sum^{n} \frac{M_i^2}{\bar{M}^2} \left(\frac{1}{m_i} - \frac{1}{M_i}\right) s_i^2 \tag{119}$$

When m_i's are equal, the variance may be calculated directly from the analysis of variance on $M_i y_{ij}/\bar{M}$, as explained in the previous section, provided S_i^2 is assumed to be constant for all i.

For N large, we have the simple expression

$$\text{Est. } V(\bar{y}_s') = \frac{s_b'^2}{n} \tag{120}$$

(*c*) *Ratio Estimate* $\bar{y}_s''$

Let

$$s_b''^2 = \frac{1}{n-1} \sum^{n} \frac{M_i^2}{\bar{M}^2} (\bar{y}_{i(m_i)} - \bar{y}_s'')^2 \tag{121}$$

On substituting for $\bar{y}_s''$ and expanding, we have

$$\bar{M}^2 (n-1) s_b''^2 = \left\{ \sum^{n} M_i^2 \bar{y}^2_{i(m_i)} - 2 \left(\sum^{n} M_i^2 \bar{y}_{i(m_i)} \right) \times \left(\frac{\sum^{n} M_i \bar{y}_{i(m_i)}}{\sum^{n} M_i} \right) + \left(\sum^{n} M_i^2 \right) \left(\frac{\sum^{n} M_i \bar{y}_{i(m_i)}}{\sum^{n} M_i} \right)^2 \right\}$$

$$= \sum^{n} M_i^2 \bar{y}^2_{i(m_i)} - \frac{2}{\sum^{n} M_i} \left\{ \sum^{n} M_i^3 \bar{y}^2_{i(m_i)} + \sum_{i \neq i'}^{n} M_i^2 M_{i'} \bar{y}_{i(m_i)} \bar{y}_{i'(m_{i'})} \right\}$$

$$+ \frac{\sum^{n} M_i^2}{\left(\sum^{n} M_i\right)^2} \left\{ \sum^{n} M_i^2 \bar{y}^2_{i(m_i)} + \sum_{i \neq i'}^{n} M_i M_{i'} \bar{y}_{i(m_i)} \bar{y}_{i'(m_{i'})} \right\}$$

Taking expectations for a fixed sample of first-stage units term by term, we obtain

$$
\begin{aligned}
\bar{M}^2(n-1)\,E(s_b''^2\,|\,i) &= \sum^{n} M_i^2\left\{\bar{y}_{i.}^2+\left(\frac{1}{m_i}-\frac{1}{M_i}\right)\mathrm{S}_i^2\right\} \\
&\quad -\frac{2}{\overset{n}{\Sigma}\,M_i}\left[\sum^{n} M_i^3\left\{\bar{y}_{i.}^2+\left(\frac{1}{m_i}-\frac{1}{M_i}\right)\mathrm{S}_i^2\right\}\right. \\
&\quad \left.+\sum_{i\neq i'}^{n} M_i^2 M_{i'}\bar{y}_{i.}\bar{y}_{i'.}\right]+\frac{\overset{n}{\Sigma}\,M_i^2}{\left(\overset{n}{\Sigma}\,M_i\right)^2}\left[\sum^{n} M_i^2\left\{\bar{y}_{i.}^2\right.\right. \\
&\quad \left.\left.+\left(\frac{1}{m_i}-\frac{1}{M_i}\right)\mathrm{S}_i^2\right\}+\sum_{i\neq i'}^{n} M_i M_{i'}\bar{y}_{i.}\bar{y}_{i'.}\right] \\
&= \sum^{n} M_i^2\left\{\bar{y}_{i.}^2+\left(\frac{1}{m_i}-\frac{1}{M_i}\right)\mathrm{S}_i^2\right\} \\
&\quad -2\left(\overset{n}{\Sigma}\,M_i^2\bar{y}_{i.}\right)\left(\frac{\overset{n}{\Sigma}\,M_i\bar{y}_{i.}}{\overset{n}{\Sigma}\,M_i}\right) \\
&\quad -\frac{2}{\overset{n}{\Sigma}\,M_i}\sum^{n} M_i^3\left(\frac{1}{m_i}-\frac{1}{M_i}\right)\mathrm{S}_i^2 \\
&\quad +\frac{\overset{n}{\Sigma}\,M_i^2}{\left(\overset{n}{\Sigma}\,M_i\right)^2}\left(\overset{n}{\Sigma}\,M_i\bar{y}_{i.}\right)^2 \\
&\quad +\frac{\overset{n}{\Sigma}\,M_i^2}{\left(\overset{n}{\Sigma}\,M_i\right)^2}\sum^{n} M_i^2 \\
&\quad \times\left(\frac{1}{m_i}-\frac{1}{M_i}\right)\mathrm{S}_i^2
\end{aligned}
$$

Combining the terms in y's and those in $\mathbf{S}_i^2$'s separately and putting

$$\bar{y}_{n.}'' = \frac{\sum^{n} M_i \bar{y}_{i.}}{\sum^{n} M_i}$$

we have

$$E(s_b''^2 \mid i) = \frac{1}{n-1}\left\{\sum^{n} \frac{M_i^2}{\bar{M}^2} (\bar{y}_{i.} - \bar{y}_{n.}'')^2 + \sum^{n} \frac{M_i^2}{\bar{M}^2}\left(\frac{1}{m_i} - \frac{1}{M_i}\right) \mathbf{S}_i^2 \left(1 - \frac{2M_i}{\sum^{n} M_i} + \frac{\sum^{n} M_i^2}{\left(\sum^{n} M_i\right)^2}\right)\right\} \tag{122}$$

Taking further expectation over first-stage samples of n and using the result (43) from Chapter IV, namely, that in large samples,

$$E\left\{\frac{1}{n-1} \sum^{n} \frac{M_i^2}{\bar{M}^2} (\bar{y}_{i.} - \bar{y}_{n.}'')^2\right\} = \mathbf{S}_b''^2$$

we obtain

$$E(s_b''^2) = \mathbf{S}_b''^2 + \frac{1}{n-1} E\left[\sum^{n} \frac{M_i^2}{\bar{M}^2}\left(\frac{1}{m_i} - \frac{1}{M_i}\right) \mathbf{S}_i^2 \times \left\{1 - \frac{2M_i}{\sum^{n} M_i} + \frac{\sum^{n} M_i^2}{\left(\sum^{n} M_i\right)^2}\right\}\right] \tag{123}$$

whence

$$\text{Est. } \mathbf{S}_b''^2 = s_b''^2 - \frac{1}{n-1} \sum^{n}\left\{\frac{M_i^2}{\bar{M}^2}\left(\frac{1}{m_i} - \frac{1}{M_i}\right) s_i^2\right\} \times \left\{1 - \frac{2M_i}{\sum^{n} M_i} + \frac{\sum^{n} M_i^2}{\left(\sum^{n} M_i\right)^2}\right\} \tag{124}$$

On substituting from (117) and (124) in (96), we get

$$\text{Est. } V(\bar{y}_s'') = \left(\frac{1}{n} - \frac{1}{N}\right) s_b''^2 + \sum^{n} \left\{\frac{M_i^2}{\bar{M}^2}\left(\frac{1}{m_i} - \frac{1}{M_i}\right) s_i^2\right\}$$

$$\times \left\{\frac{1}{n^2} - \frac{1}{n-1}\left(\frac{1}{n} - \frac{1}{N}\right)\right.$$

$$\left.\times \left(1 - \frac{2M_i}{\overset{n}{\Sigma} M_i} + \frac{\overset{n}{\Sigma} M_i^2}{\left(\overset{n}{\Sigma} M_i\right)^2}\right)\right\}$$

or, to a first approximation,

$$= \left(\frac{1}{n} - \frac{1}{N}\right) s_b''^2 + \frac{1}{nN} \sum^{n} \frac{M_i^2}{\bar{M}^2}\left(\frac{1}{m_i} - \frac{1}{M_i}\right) s_i^2 \tag{125}$$

(d) Ratio Estimate $\bar{y}_R$

The steps leading to the estimation from the sample of the variance of $\bar{y}_R$ are similar to those given in (*c*) above. We shall quote here only the final result. We write, to a first approximation,

$$\text{Est. } V_1(\bar{y}_R) = \left(\frac{1}{n} - \frac{1}{N}\right) \frac{1}{n-1} \sum^{n} u_i^2 (\bar{y}_{i(m_i)} - R_s \bar{x}_{i(m_i)})^2$$

$$+ \frac{1}{nN} \sum^{n} u_i^2 \left(\frac{1}{m_i} - \frac{1}{M_i}\right) d_i^2 \tag{126}$$

where

$$d_i^2 = s_{iy}^2 + R_s^2 s_{ix}^2 - 2R_s s_{iyx} \tag{127}$$

7.14 The Use of the Analysis of Variance Method to Compute the Variance of the Sample Mean

It has been pointed out already that unless m_i's are equal it is not valid to evaluate the variance of the estimate using the analysis of variance table. Nevertheless, for moderate inequality in m_i's, the method with certain adjustments has been recommended for use (Yates, 1949). The method consists in calculating a number λ given by

$$\lambda = \frac{\sum^{n} m_i - \dfrac{\sum^{n} m_i^2}{\sum^{n} m_i}}{n-1} \tag{128}$$

and using it to estimate S_b^2 by means of

$$\text{Est. } S_b^2 = \frac{B - W}{\lambda} \tag{129}$$

where B and W denote the mean squares in the analysis of variance table for the sample. On account of its simplicity the method makes an appeal to the practical worker. It is therefore important to examine the conditions when it can be used.

Let us suppose that

(*a*) N, the number of first-stage units in the population, is large;

(*b*) M_i, the number of second-stage units in each first-stage unit, is equal to M;

(*c*) m_i is the number of second-stage units to be drawn from the first-stage unit selected at the i-th draw ($i=1, 2, \ldots, n$); and

(*d*) S_i^2 is constant for all i and equal to, say, S_w^2.

Consider the estimate

$$\bar{y}_{m_0} = \frac{1}{m_0} \sum_{i}^{n} \sum_{j}^{m_i} y_{ij} \tag{130}$$

where m_0 has the usual meaning $\sum^{n} m_i$. It is easy to see that this is an unbiased estimate of the population mean $\bar{y}_{..}$. For,

$$\begin{aligned}
E\,(\bar{y}_{m_0}) &= \frac{1}{m_0}\,E\left\{\sum^{n} m_i E\,(\bar{y}_{i(m_i)} \mid i)\right\} \\
&= \frac{1}{m_0}\,E\left\{\sum^{n} m_i \bar{y}_{i.}\right\} \\
&= \frac{1}{m_0}\sum^{n} m_i E\,(\bar{y}_{i.}) \\
&= \bar{y}_{..}
\end{aligned} \tag{131}$$

The variance of $\bar{y}_{m_0}$ is given by

$$\begin{aligned}
V\,(\bar{y}_{m_0}) &= E\,(\bar{y}_{m_0} - \bar{y}_{..})^2 \\
&= \frac{1}{m_0^2}\,E\left\{\sum^{n} m_i\,(\bar{y}_{i(m_i)} - \bar{y}_{..})\right\}^2 \\
&= \frac{1}{m_0^2}\,E\left\{\sum^{n} m_i^2\,(\bar{y}_{i(m_i)} - \bar{y}_{..})^2 \right.\\
&\qquad \left. + \sum^{n}_{i \neq i'} m_i m_{i'}\,(\bar{y}_{i(m_i)} - \bar{y}_{..})\,(\bar{y}_{i'(m_{i'})} - \bar{y}_{..})\right\} \\
&= \frac{1}{m_0^2}\,E\left\{\sum^{n} m_i^2\,(\bar{y}_{i(m_i)} - \bar{y}_{i.} + \bar{y}_{i.} - \bar{y}_{..})^2 \right.\\
&\qquad + \sum^{n}_{i \neq i'} m_i m_{i'}\,(\bar{y}_{i(m_i)} - \bar{y}_{i.} + \bar{y}_{i.} - \bar{y}_{..}) \\
&\qquad \left. \times (\bar{y}_{i'(m_{i'})} - \bar{y}_{i'.} + \bar{y}_{i'.} - \bar{y}_{..})\right\} \\
&= \frac{1}{m_0^2}\,E\left[\sum^{n} m_i^2\,\{(\bar{y}_{i(m_i)} - \bar{y}_{i.})^2 + (\bar{y}_{i.} - \bar{y}_{..})^2 \right.\\
&\qquad + 2\,(\bar{y}_{i(m_i)} - \bar{y}_{i.})\,(\bar{y}_{i.} - \bar{y}_{..})\} \\
&\qquad + \sum^{n}_{i \neq i'} m_i m_{i'}\ \{(\bar{y}_{i(m_i)} - \bar{y}_{i.})\ (\bar{y}_{i'(m_{i'})} - \bar{y}_{i'.}) \\
&\qquad + (\bar{y}_{i(m_i)} - \bar{y}_{i.})\ (\bar{y}_{i'.} - \bar{y}_{..}) \\
&\qquad + (\bar{y}_{i.} - \bar{y}_{..})\,(\bar{y}_{i'(m_{i'})} - \bar{y}_{i'.}) \\
&\qquad \left. + (\bar{y}_{i.} - \bar{y}_{..})\,(\bar{y}_{i'.} - \bar{y}_{..})\}\right]
\end{aligned}$$

$$= \frac{1}{m_0^2}\left[\sum^n m_i^2 \{E(\bar{y}_{i(m_i)} - \bar{y}_{i.})^2 + E(\bar{y}_{i.} - \bar{y}_{..})^2\}\right.$$

$$\left. + \sum_{i \neq i'}^n m_i m_{i'}\, E\{(\bar{y}_{i.} - \bar{y}_{..})(\bar{y}_{i'.} - \bar{y}_{..})\}\right]$$

since the expectations of all the other terms are clearly zero. Hence

$$V(\bar{y}_{m_0}) = \frac{1}{m_0^2}\left[\sum^n m_i^2 \left\{\left(\frac{1}{m_i} - \frac{1}{M}\right) S_w^2 + S_b^2\right\}\right.$$

$$\left. - \sum_{i \neq i'}^n m_i m_{i'} \frac{S_b^2}{N}\right]$$

$$= \frac{S_w^2}{m_0}\left\{1 - \frac{\sum^n m_i^2}{m_0 M}\right\} + \frac{\sum^n m_i^2}{m_0^2}\, S_b^2 \qquad (132)$$

neglecting terms containing $1/N$. To evaluate S_b^2, we start with *B*, the mean square in the analysis of variance for the sample and take expectations. We obtain

$$E(B) = E\left\{\frac{\sum^n m_i (\bar{y}_{i(m_i)} - \bar{y}_{m_0})^2}{n-1}\right\}$$

$$= \frac{1}{n-1}\, E\left\{\sum^n m_i \bar{y}^2_{i(m_i)} - m_0 \bar{y}_{m_0}^2\right\}$$

$$= \frac{1}{n-1}\left[E \sum^n m_i \left\{\bar{y}_{i.}^2 + S_w^2\left(\frac{1}{m_i} - \frac{1}{M}\right)\right\}\right.$$

$$\left. - m_0\{\bar{y}_{..}^2 + V(\bar{y}_{m_0})\}\right]$$

$$= \frac{1}{n-1}\left[S_b^2\left(m_0 - \frac{\sum^n m_i^2}{m_0}\right) + \overline{n-1}\, S_w^2\right.$$

$$\left. - \frac{S_w^2}{M}\left(m_0 - \frac{\sum^n m_i^2}{m_0}\right)\right]$$

$$= \lambda S_b^2 + S_w^2\left(1 - \frac{\lambda}{M}\right) \qquad (133)$$

whence

$$\text{Est. } S_b^2 = \frac{B - W\left(1 - \frac{\lambda}{M}\right)}{\lambda} \tag{134}$$

For M large, we have

$$\text{Est. } S_b^2 = \frac{B - W}{\lambda} \tag{135}$$

We may infer that if the sampling fraction at each stage is small and the variation in the size of first-stage units is negligible, the method may give reliable results. Its use under conditions other than those specified above will need to be justified. It should also be mentioned that the system of drawing m_i second-stage units at the i-th draw irrespective of which first-stage unit is included in the sample at the i-th draw is not a rational system which is likely to be used in practice. In an efficient survey design, m_i's will be usually equal within a stratum, although it is likely that through extraneous causes the numbers of second-stage units actually collected in the sample may be unequal. If these extraneous causes are random causes, in other words, if m_1, m_2, . . ., m_n can be considered a random sample from the respective first-stage units and further the variation among M_i's is not large, this method as in (134), may give a sufficiently good estimate of the variance when N is large.

7.15 Allocation of Sample

In our discussion so far, we have assumed that the number of second-stage units to be drawn from the i-th first-stage unit, namely m_i, is any arbitrary number less than M_i. It may be related in some way to the size of the i-th unit, as for example, when it is proportional to M_i, or it may be independent of it. The guiding principle in choosing it is clearly to maximize the precision of the estimate for given cost or minimize the cost for desired precision.

We shall suppose that the total cost consists of two components, one depending upon the number of first-stage units in the sample c_1n, and the other on the total number of second-stage units in the sample, namely $c_2 \overset{n}{\Sigma}\, m_i$. This second component will, however, vary from sample to sample of n first-stage units. We shall therefore consider the average cost instead of the actual cost of surveying a sample, given by

$$C = c_1 n + c_2 \frac{n}{N} \sum_{i=1}^{N} m_i \tag{136}$$

and proceed to determine the optimum allocation. We shall suppose that the estimate to be used is the unbiased estimate $\bar{y}_s{}'$.

Now, from (91) and (136), we obtain

$$\left(V(\bar{y}_s{}') + \frac{S_b'^2}{N}\right) C = \left\{ S_b'^2 - \frac{1}{N\bar{M}} \sum_{i=1}^{N} \left(\frac{1}{\bar{M}}\, M_i S_i^2\right) + \frac{1}{N} \sum_{i=1}^{N} \frac{M_i^2}{\bar{M}^2} \cdot \frac{S_i^2}{m_i} \right\} \left(c_1 + \frac{c_2}{N} \sum_{i=1}^{N} m_i\right) \tag{137}$$

$$= c_1 \Delta + \frac{c_2}{N^2} \sum_{i=1}^{N} \frac{M_i^2}{\bar{M}^2} S_i^2 + \frac{1}{N} \sum_{i=1}^{N} \left(c_2 \Delta m_i + c_1 \frac{M_i^2}{\bar{M}^2} \cdot \frac{S_i^2}{m_i}\right) + \frac{c_2}{N^2 \bar{M}^2} \sum_{i>i'=1}^{N} \left(\frac{m_i}{m_{i'}}\, M_{i'}^2 S_{i'}^2 + \frac{m_{i'}}{m_i}\, M_i^2 S_i^2\right) \tag{138}$$

where

$$\Delta = S_b'^2 - \frac{1}{N\bar{M}} \sum_{i=1}^{N} M_i \frac{S_i^2}{\bar{M}} \tag{139}$$

Assuming Δ to be positive, the right-hand side of (138) can be put in the form

$$c_1\Delta + \frac{c_2}{N^2}\sum_{i=1}^{N}\frac{M_i^2}{\bar{M}^2}\mathbf{S}_i^2 + \frac{2}{N\bar{M}}\sum_{i=1}^{N}\sqrt{c_1c_2\Delta}\,M_i\mathbf{S}_i$$

$$+ \frac{2c_2}{N^2\bar{M}^2}\sum_{i>i'=1}^{N}M_iM_{i'}\mathbf{S}_i\mathbf{S}_{i'}$$

$$+ \frac{1}{N}\sum_{i=1}^{N}\left(\sqrt{c_2\Delta m_i} - \sqrt{\frac{c_1}{m_i}}\cdot\frac{M_i\mathbf{S}_i}{\bar{M}}\right)^2$$

$$+ \frac{c_2}{N^2\bar{M}^2}\sum_{i>i'=1}^{N}\left(\sqrt{\frac{m_i}{m_{i'}}}\,M_{i'}\mathbf{S}_{i'} - \sqrt{\frac{m_{i'}}{m_i}}\,M_i\mathbf{S}_i\right)^2 \qquad (140)$$

and this is minimum when each of the two square terms is zero, giving us

$$m_i = \sqrt{\frac{c_1}{c_2\Delta}}\cdot\frac{M_i}{\bar{M}}\cdot\mathbf{S}_i \qquad (i = 1, 2, \ldots, N) \qquad (141)$$

We notice that m_i is proportional to the product of three factors: the first depending upon the cost factors, the second upon the size of the selected unit and the third on the variation of the character under study within the selected unit. Advance knowledge of $\mathbf{S}_i^2$ is, however, difficult to obtain. Practical considerations require that m_i should be independent of $\mathbf{S}_i^2$, even if this means departing somewhat from the optimum. One method of choosing m_i is to make it proportional to M_i. This would imply the assumption that $\mathbf{S}_i^2$ is constant for all i. Usually $\mathbf{S}_i^2$ will be found to increase with M_i, although perhaps not as fast as M_i. One method of reducing the dependence of m_i on $\mathbf{S}_i^2$ is to group together into strata first-stage units of about the same size, provided stratification by size does not prevent stratification by other and more important characters, and choose m_i proportional to M_i within the several strata.

Suppose then m_i is so determined as to be proportional to M_i, say,

$$m_i = kM_i \tag{142}$$

where k is some positive constant. Substituting from (142) in (137) for m_i, we get

$$\begin{aligned}\left(V(\bar{y}_s') + \frac{\mathbf{S}_b'^2}{N}\right)C &= \left\{\Delta + \frac{1}{N\bar{M}^2}\cdot\frac{1}{k}\sum_{i=1}^{N} M_i\mathbf{S}_i^2\right\}(c_1+c_2k\bar{M})\\ &= \Delta c_1 + \frac{c_2}{N\bar{M}}\sum_{i=1}^{N} M_i\mathbf{S}_i^2\\ &\quad + 2\sqrt{\frac{c_1c_2\Delta}{N\bar{M}}\sum_{i=1}^{N} M_i\mathbf{S}_i^2}\\ &\quad + \left(\sqrt{\frac{c_1}{N\bar{M}^2}\cdot\frac{1}{k}\sum_{i=1}^{N} M_i\mathbf{S}_i^2} - \sqrt{\Delta c_2\bar{M}k}\right)^2\end{aligned}$$

which is minimum for (143)

$$k = \sqrt{\frac{c_1}{c_2}\cdot\frac{1}{N\bar{M}^3\Delta}\sum_{i=1}^{N} M_i\mathbf{S}_i^2} \tag{144}$$

By analogy with (39) this may be approximated by

$$\sqrt{\frac{c_1}{c_2}\cdot\frac{1}{\bar{M}^2}\cdot\frac{1-\bar{\rho}}{\bar{\rho}}} \tag{145}$$

where $\bar{\rho}$ may be termed the average intra-class correlation over all units in the stratum. It follows that k is determined by the same considerations as those discussed earlier for the case of equal clusters. Knowing k, the value of $\hat{n}$ is obtained from (91) or (136) according as V or C is fixed in advance and C or V is minimized.

The reader may verify that the optimum allocation is governed by the same formulæ as those presented here even when any of the other consistent estimates is used. Thus, for the estimate

$\bar{y}_s''$, he will notice that the sampling variance has the same form as that for the estimate $\bar{y}_s'$ except that $S_b'^2$ is replaced by $S_b''^2$. It follows that the optimum value of m_i is so determined that

$$m_i = \sqrt{\frac{c_1}{c_2 \Delta'}} \cdot \frac{M_i}{\bar{M}} S_i$$

where

$$\Delta' = S_b''^2 - \frac{1}{N\bar{M}} \sum_{i=1}^{N} \frac{M_i}{\bar{M}} S_i^2$$

For S_i constant, we obtain

$$m_i = k' M_i$$

where k' is a positive constant.

Example 7.2

A corn borer survey is carried out every fall in Iowa (U.S.A.) for estimating the number of borers per plant. Fifty sampling units of 25 stalks are selected at random from each district, and for each sampling unit the number of corn borers per stalk is estimated. Since it is costly to dissect all of the infested plants in each sampling unit, a sub-sample of two is dissected to obtain the estimate of the number of corn borers per infested plant. When the number of infested plants in a sampling unit is one, that is dissected.

Two methods of estimation are followed:

(*a*) The first method consists in computing the simple mean of the number of borers per plant from each sampling unit. Thus, if M_i is the number of infested plants in the i-th sampling unit and $\bar{y}_{i(m_i)}$ the estimated number of borers per infested plant in the i-th sampling unit and n the number of units in the sample, then an estimate of μ, the number of borers per 25 plants, is given by

$$z_1 = \frac{1}{n} \overset{n}{\Sigma} M_i \bar{y}_{i(m_i)}$$

(*b*) The other method of estimation is to compute z_2, where

$$z_2 = \frac{\overset{n}{\Sigma} M_i}{n} \cdot \frac{\overset{n}{\Sigma} \bar{y}_{i(m_i)}}{n} = \bar{M}_n \bar{y}_{n(m_i)}$$

i.e., the product of the average infestation per sampling unit and the average number of borers per infested plant per sampling unit.

Columns 2, 3 and 4 of Table 7.6 contain the relevant data for one district. Obtain the two estimates. Examine whether they are unbiased estimates of the population value μ for the district. Also calculate the mean square errors of the two estimates.

The estimate z_1 corresponds to the estimate $\bar{y}_s'$ in the text. It is therefore an unbiased estimate of the population value. In cols. 2, 5 and 6 of Table 7.6 are given the values of M_i, $\bar{y}_{i(m_i)}$ and of the products $M_i\bar{y}_{i(m_i)}$, called g_i, and the means of all the three for the 50 units in the sample. We find that

$$z_1 = \frac{1}{n} \overset{n}{\Sigma}\, g_i = 12{\cdot}4$$

The variance of z_1 can be directly calculated from (120) after putting $\bar{M} = 1$ in that expression. We obtain

$$\text{Est. } V(z_1) = \frac{1}{n}\, s_g^2$$

$$= \frac{1}{n(n-1)} \left\{ \overset{n}{\Sigma}\, g_i^2 - n\bar{g}^2 \right\}$$

$$= \frac{19463 - 7750}{50 \times 49}$$

$$= 4{\cdot}78$$

Turning to the second estimate, on substituting from Table 7.6 the values of $\bar{M}_n$ and $\bar{y}_{n(m_i)}$ in the expression for z_2, we obtain

$$z_2 = 13{\cdot}34 \times 0{\cdot}80 = 10{\cdot}7$$

To obtain the expected value of z_2, we have

$$E(z_2) = \frac{1}{n^2} E\left\{\left(\sum^n M_i\right)\left(\sum^n E(\bar{y}_{i(m_i)} \mid i)\right)\right\}$$

$$= \frac{1}{n^2} E\left\{\left(\sum^n M_i\right)\left(\sum^n \bar{y}_{i.}\right)\right\}$$

$$= \frac{1}{n^2} E\left\{\sum^n M_i \bar{y}_{i.} + \sum_{i \neq i'}^n M_i \bar{y}_{i'.}\right\}$$

$$= \frac{1}{n}\left\{\mu + (n-1)\left(\frac{N}{N-1}\bar{M}\bar{y}_{N.} - \frac{\mu}{N-1}\right)\right\}$$

$$= \frac{1}{n}\{\mu + (n-1)\,\bar{M}\bar{y}_{N.}\}$$

for large N.

Define

$$\rho = \frac{E(M_i - \bar{M})(\bar{y}_{i.} - \bar{y}_{N.})}{\sqrt{E(M_i - \bar{M})^2 E(\bar{y}_{i.} - \bar{y}_{N.})^2}}$$

Then we can put

$$E(z_2) = \mu - \frac{n-1}{n} \cdot \frac{N-1}{N} \rho \mathbf{S}_M \mathbf{S}_b$$

where

$$\mathbf{S}_M{}^2 = \frac{1}{N-1} \sum_{i=1}^{N} (M_i - \bar{M})^2$$

and

$$\mathbf{S}_b{}^2 = \frac{1}{N-1} \sum_{i=1}^{N} (\bar{y}_{i.} - \bar{y}_{N.})^2$$

or

$$E(z_2) = \mu - \frac{n-1}{n} \rho \mathbf{S}_M \mathbf{S}_b$$

since N is large. We note that z_2 is a biased estimate of μ. To evaluate the bias we need the estimates of ρ, $\mathbf{S}_M$ and $\mathbf{S}_b$. These can be taken directly from Table 7.6. We have

TABLE 7.6

Corn Borer Survey in Iowa, 1951

Serial No. of Sampling Unit	No. of Infested Plants M_i	No. of Borers on the Infested Plants		$\bar{y}_{i(m_i)}$	$M_i\bar{y}_{i(m_i)} = g_i$	$(y_{i1}-y_{i2})^2/2 = s_i^2$	$\left(\frac{1}{m_i}-\frac{1}{M_i}\right)$	$\left(\frac{1}{m_i}-\frac{1}{M_i}\right) s_i^2$
		y_{i1}	y_{i2}					
1	14	1	0	0·5	7·0	0·5	0·429	0·214
2	9	2	2	2·0	18·0	0	..	0
3	0	..	..	0	0	0	..	0
4	1	0	..	0	0	0	..	0
5	1	0	..	0	0	0	..	0
6	2	0	0	0	0	0	..	0
7	13	0	0	0	0	0	..	0
8	0	..	..	0	0	0	..	0
9	3	1	1	1·0	3·0	0	..	0
10	3	0	0	0	0	0	..	0
11	9	0	0	0	0	0	..	0
12	11	4	3	3·5	38·5	0·5	0·409	0·204
13	8	1	1	1·0	8·0	0	..	0
14	3	0	1	0·5	1·5	0·5	0·167	0·084
15	13	2	1	1·5	19·5	0·5	0·423	0·212
16	22	0	2	1·0	22·0	2·0	0·455	0·910
17	21	2	0	1·0	21·0	2·0	0·452	0·904
18	5	2	0	1·0	5·0	2·0	0·300	0·600
19	21	1	1	1·0	21·0	0	..	0
20	11	0	0	0	0	0	..	0
21	2	1	0	0·5	1·0	0·5	0	0
22	12	0	0	0	0	0	..	0
23	24	0	0	0	0	0	..	0
24	13	0	1	0·5	6·5	0·5	0·423	0·212
25	21	0	1	0·5	10·5	0·5	0·452	0·226
26	23	3	0	1·5	34·5	4·5	0·457	2·056
27	17	2	0	1·0	17·0	2·0	0·441	0·882
28	24	2	0	1·0	24·0	2·0	0·458	0·916
29	18	4	1	2·5	45·0	4·5	0·444	1·998
30	21	1	1	1·0	21·0	0	..	0
31	12	2	1	1·5	18·0	0·5	0·417	0·208

TABLE 7.6—*Contd.*

Serial No. of Sampling Unit	No. of Infested Plants M_i	No. of Borers on the Infested Plants		$\bar{y}_{i(m_i)}$	$M_i\,\bar{y}_{i(m_i)} = g_i$	$(y_{i1}-y_{i2})^2/2 = s_i^2$	$\left(\frac{1}{m_i} - \frac{1}{M_i}\right)$	$\left(\frac{1}{m_i} - \frac{1}{M_i}\right) s_i^2$
		y_{i1}	y_{i2}					
32	11	2	0	1·0	11·0	2·0	0·409	0·818
33	16	0	0	0	0	0	..	0
34	23	2	2	2·0	46·0	0	..	0
35	17	2	5	3·5	59·5	4·5	0·441	1·984
36	10	1	1	1·0	10·0	0	..	0
37	18	1	0	0·5	9·0	0·5	0·444	0·222
38	19	0	0	0	0	0	..	0
39	17	0	1	0·5	8·5	0·5	0·441	0·220
40	19	1	0	0·5	9·5	0·5	0·447	0·224
41	24	1	0	0·5	12·0	0·5	0·458	0·229
42	14	1	1	1·0	14·0	0	..	0
43	25	0	1	0·5	12·5	0·5	0·460	0·230
44	21	0	0	0	0	0	..	0
45	8	1	0	0·5	4·0	0·5	0·375	0·188
46	19	0	1	0·5	9·5	0·5	0·447	0·224
47	25	3	2	2·5	62·5	0·5	0·460	0·230
48	0	..	..	0	0	0	..	0
49	1	1	..	1·0	1·0	0	..	0
50	23	0	1	0·5	11·5	0·5	0·457	0·228
Sums	667			40·0	622·5			14·42
Means	13·3			0·8	12·4			0·288
Sums of squares	12111			68·0	19463			
Mean square deviations (s_M^2, s_b^2)	65·58			0·7347				
s_M, s_b	8·10			0·857				

$$r^2 = \frac{(622\cdot5-40\times13\cdot3)^2}{(36)\,(3213)} = 0\cdot0683 \;;\quad r = 0\cdot261$$

We are indebted to Dr. T. A. Bancroft for data in this table.

$$\hat{\rho} = r = 0{\cdot}261$$

$$\hat{S}_M{}^2 = s_M{}^2 = 65{\cdot}58$$

$$\hat{S}_M = s_M = 8{\cdot}10$$

while $\hat{S}_b{}^2$ is obtained from (112), being given by

$$\begin{aligned}\hat{S}_b{}^2 &= s_b{}^2 - \frac{1}{n}\sum^{n}\left(\frac{1}{m_i} - \frac{1}{M_i}\right)s_i{}^2\\ &= 0{\cdot}7347 - \frac{1}{50}\,(14{\cdot}42)\\ &= 0{\cdot}4463\end{aligned}$$

$$\hat{S}_b = 0{\cdot}668$$

Hence

$$\begin{aligned}\text{bias} &= \frac{n-1}{n}\,\hat{\rho}\hat{S}_M\hat{S}_b\\ &= \frac{49}{50}\,(0{\cdot}261)\,(8{\cdot}10)\,(0{\cdot}668)\\ &= 1{\cdot}38\end{aligned}$$

The derivation of the variance of z_2 is rather complex, but for large N and n, and n/N negligible, it presents no difficulty. We write to a first approximation,

$$V(\bar{M}_n\bar{y}_{n(m_i)}) = \bar{y}_{N.}{}^2V(\bar{M}_n) + \bar{M}^2V(\bar{y}_{n(m_i)}) + 2\bar{y}_{N.}\bar{M}\,\mathrm{Cov}\,(\bar{M}_n,\ \bar{y}_{n(m_i)})$$

whence to terms in $1/n$, we get

$$\text{Est. } V(\bar{M}_n\bar{y}_{n(m_i)}) = \bar{y}^2{}_{n(m_i)}\frac{s_M{}^2}{n} + \frac{\bar{M}_n{}^2}{n}\,s_b{}^2 + 2\bar{y}_{n(m_i)}\bar{M}_n\,\frac{rs_Ms_b}{n}$$

On substituting from Table 7.6, we get

$$\begin{aligned}\text{Est. } V(z_2) &= \frac{(0{\cdot}64)\,(65{\cdot}58)}{50} + \frac{(178)\,(0{\cdot}7347)}{50}\\ &\quad + \frac{2\,(0{\cdot}8)\,(13{\cdot}34)\,(0{\cdot}261)\,(8{\cdot}10)\,(0{\cdot}857)}{50}\\ &= 0{\cdot}839 + 2{\cdot}616 + 0{\cdot}773\\ &= 4{\cdot}23\end{aligned}$$

and

$$\begin{aligned} M.S.E.(z_2) &= V(z_2) + \text{bias}^2 \\ &= 4{\cdot}23 + 1{\cdot}90 \\ &= 6{\cdot}13 \end{aligned}$$

7.16 Stratified Sub-Sampling

By far the most common design in surveys is stratified multistage sampling. In this design the population of first-stage units is first divided into strata, within each stratum a sample of first-stage units is selected and each of the selected first-stage units is further sub-sampled. Crop surveys with the subdivision as the stratum, described in Example 7.1 and the corn borer survey with the district as the stratum described in Example 7.2, are examples of this design. In this section we shall give the formulæ for the estimate of the population mean in stratified two-stage sampling, and its variance. We shall consider the unbiased estimate only.

Let the population be divided into k strata with N_t first-stage units in the t-th stratum, so that

$$\sum_{t=1}^{k} N_t = N$$

Further, we shall denote by M_{ti} the number of second-stage units in the i-th first-stage unit of the t-th stratum, with M_{to} denoting the total number of second-stage units in the t-th stratum, *i.e.*,

$$M_{to} = \sum_{i=1}^{N_t} M_{ti} = N_t \bar{M}_t$$

Let n_t denote the number of first-stage units to be included in the sample from the t-th stratum, so that

$$n = \sum_{t=1}^{k} n_t$$

and m_{ti} denote the number of second-stage units to be sampled from the i-th selected first-stage unit.

Following the previous notation, we shall denote by

$\bar{y}_{t..}$ = the population mean per second-stage unit in the *t*-th stratum

$$= \frac{1}{M_{to}} \sum_{i=1}^{N_t} \sum_{j=1}^{M_{ti}} y_{tij}$$

$\bar{y}_{t\cdot}{}'$ = the corresponding sample mean for the *t*-th stratum

$$= \frac{1}{\bar{M}_t n_t} \sum_{i}^{nt} \frac{M_{ti}}{m_{ti}} \sum_{j}^{m_{ti}} y_{tij}$$

$\bar{y}_{..}$ = the population mean per second-stage unit

$$= \frac{\sum\limits_{t=1}^{k} M_{to}\, \bar{y}_{t.}}{\sum\limits_{t=1}^{k} M_{to}}$$

$$= \sum_{t=1}^{k} \lambda_t\, \bar{y}_{t..}$$

where

$$\lambda_t = \frac{M_{to}}{M_0} = \frac{\sum\limits_{i=1}^{N_t} M_{ti}}{\sum\limits_{t=1}^{k} \sum\limits_{i=1}^{N_t} M_{ti}}$$

and

$\bar{y}_w$ = the corresponding sample estimate

$$= \sum_{t=1}^{k} \lambda_t\, \bar{y}_{t\cdot}{}' \tag{146}$$

Clearly, $\bar{y}_w$ is an unbiased estimate of the population mean, while its variance is given by

$$V(\bar{y}_w)_S = \sum_{t=1}^{k} \lambda_t^2\, V(\bar{y}_{t\cdot}{}')$$

Substituting from (91), we have

$$V(\bar{y}_w)_S = \sum_{t=1}^{k} \lambda_t^2 \left\{ \left(\frac{1}{n_t} - \frac{1}{N_t} \right) \mathbf{S}_{tb}'^2 + \frac{1}{n_t N_t} \sum_{i=1}^{N_t} \frac{M_{ti}^2}{\bar{M}_t^2} \left(\frac{1}{m_{ti}} - \frac{1}{M_{ti}} \right) \mathbf{S}_{ti}^2 \right\} \tag{147}$$

where

$$\left. \begin{aligned} \mathbf{S}_{tb}'^2 &= \frac{1}{N_t - 1} \sum_{i=1}^{N_t} \left(\frac{M_{ti}}{\bar{M}_t} \bar{y}_{ti.} - \bar{y}_{t..} \right)^2 \\ &\text{and} \\ \mathbf{S}_{ti}^2 &= \frac{1}{(M_{ti} - 1)} \sum_{j=1}^{M_{ti}} (y_{tij} - \bar{y}_{ti.})^2 \end{aligned} \right\} \tag{148}$$

and estimates of $\mathbf{S}_{tb}'^2$ and $\mathbf{S}_{ti}^2$ are provided by

$$\left. \begin{aligned} \text{Est. } \mathbf{S}_{tb}'^2 &= s_{tb}'^2 - \frac{1}{n_t} \sum^{n_t} \frac{M_{ti}^2}{\bar{M}_t^2} \left(\frac{1}{m_{ti}} - \frac{1}{M_{ti}} \right) s_{ti}^2 \\ \text{Est. } \mathbf{S}_{ti}^2 &= s_{ti}^2 = \frac{1}{m_{ti} - 1} \sum_{j}^{m_{ti}} \left(y_{tij} - \bar{y}_{ti(m_{ti})} \right)^2 \end{aligned} \right\} \tag{149}$$

Formulæ for the mean and its variance in stratified sampling appropriate for the case of equal clusters follow as special cases. Thus, for $M_{ti} = M_t$ and $m_{ti} = m_t$, we have

$$\bar{y}_w = \sum_{t=1}^{k} \lambda_t \bar{y}_{tn_tm_t}$$

where

$$\bar{y}_{tn_tm_t} = \frac{1}{n_t m_t} \sum_{i}^{n_t} \sum_{j}^{m_t} y_{tij} = \bar{y}_{ts}$$

and

$$\lambda_t = \frac{N_t M_t}{\sum_{t=1}^{k} N_t M_t}$$

and for its variance

$$V(\bar{y}_w)_S = \sum_{t=1}^{k} \lambda_t^2 \left\{ \left(\frac{1}{n_t} - \frac{1}{N_t}\right) \mathbf{S}_{tb}^2 + \frac{1}{n_t}\left(\frac{1}{m_t} - \frac{1}{M_t}\right) \bar{\mathbf{S}}_{tw}^2 \right\}$$

where

$$\mathbf{S}_{tb}^2 = \frac{1}{N_t - 1} \sum_{i=1}^{N_t} (\bar{y}_{ti.} - \bar{y}_{t..})^2$$

and

$$\bar{\mathbf{S}}_{tw}^2 = \frac{1}{N_t} \sum_{i=1}^{N_t} \frac{1}{(M_t - 1)} \sum_{j=1}^{M_t} (y_{tij} - \bar{y}_{ti.})^2$$

and estimates of $\mathbf{S}_{tb}^2$ and $\bar{\mathbf{S}}_{tw}^2$ are provided by the same formulæ as in (110) and (112), namely,

$$\left.\begin{aligned} \text{Est. } \mathbf{S}_{tb}^2 &= s_{tb}^2 - \left(\frac{1}{m_t} - \frac{1}{M_t}\right) \bar{s}_{tw}^2 \\ \text{Est. } \bar{\mathbf{S}}_{tw}^2 &= \bar{s}_{tw}^2 \end{aligned}\right\} \tag{150}$$

so that

$$\text{Est. } V(\bar{y}_w)_S = \sum_{t=1}^{k} \lambda_t^2 \left\{ \left(\frac{1}{n_t} - \frac{1}{N_t}\right) s_{tb}^2 + \frac{1}{N_t}\left(\frac{1}{m_t} - \frac{1}{M_t}\right) \bar{s}_{tw}^2 \right\} \tag{151}$$

7.17 Efficiency of Stratification in Sub-Sampling

We shall consider the simple case for which $M_{ti} = M$. Further, we shall suppose that $m_{ti} = m$, so that the total number of second-stage units to be included in the sample is a fixed number nm whether drawn as a stratified or unstratified sample. In this section we shall estimate from a given stratified sample, the difference between the variances of a stratified and an unstratified sample.

If the sample were selected as an unstratified two-stage sample, the estimate of the population mean would be

$$\bar{y}_{nm} = \frac{1}{nm} \sum_{i}^{n} \sum_{j}^{m} y_{ij} \tag{152}$$

with the sampling variance given by

$$V(\bar{y}_{nm})_{US} = \left(\frac{1}{n} - \frac{1}{N}\right) \mathrm{S}_b^2 + \left(\frac{1}{m} - \frac{1}{M}\right) \frac{\bar{\mathrm{S}}_w^2}{n} \tag{153}$$

where the letters "*US*" stand for "unstratified",

S_b^2 = the mean square between the first-stage unit means in the population

$$= \frac{1}{N-1} \sum_{i=1}^{N} (\bar{y}_{i.} - \bar{y}_{..})^2 \tag{154}$$

and

$\bar{\mathrm{S}}_w^2$ = the mean square between second-stage units within first-stage units in the whole population

$$= \frac{1}{N\ (M\ -\ 1)} \sum_{i=1}^{N} \sum_{j=1}^{M} (y_{ij} - \bar{y}_{i.})^2 \tag{155}$$

If the sample is a stratified sample, the estimate will be

$$\bar{y}_w = \sum_{t=1}^{k} p_t \bar{y}_{ts}$$

where p_t is the weight for the t-th stratum, and its sampling variance

$$V(\bar{y}_w)_S = \sum_{t=1}^{k} p_t^2 \left\{ \left(\frac{1}{n_t} - \frac{1}{N_t}\right) \mathrm{S}_{tb}^2 + \frac{1}{n_t}\left(\frac{1}{m} - \frac{1}{M}\right) \bar{\mathrm{S}}_{tw}^2 \right\} \tag{156}$$

The relative excess of (153) over (156) represents the gain in precision due to stratification. For estimating this gain from the selected stratified sample, we require an estimate of S_b^2. We have

$$(N-1)\ \mathrm{S}_b^2 = \sum_{i=1}^{N} (\bar{y}_{i.} - \bar{y}_{..})^2$$

$$= \sum_{t=1}^{k} \sum_{i=1}^{N_t} (\bar{y}_{ti.} - \bar{y}_{..})^2$$

$$= \sum_{t=1}^{k} \sum_{i=1}^{N_t} (\bar{y}_{ti.} - \bar{y}_{t..} + \bar{y}_{t..} - \bar{y}_{..})^2$$

$$= \sum_{t=1}^{k} (N_t - 1)\, \mathrm{S}_{tb}^2 + \sum_{t=1}^{k} N_t (\bar{y}_{t..} - \bar{y}_{..})^2$$

$$= \sum_{t=1}^{k} (N_t - 1)\, \mathrm{S}_{tb}^2 + \sum_{t=1}^{k} N_t \bar{y}_{t..}^2 - N\bar{y}_{..}^2 \qquad (157)$$

The estimate of S_{tb}^2 is known from (150), so that our problem reduces to estimating the second and third terms in (157). Now, from (10), we have

$$V(\bar{y}_{tn_tm}) = V(\bar{y}_{ts}) = E(\bar{y}_{ts}^2) - \bar{y}_{t..}^2$$

$$= \left(\frac{1}{n_t} - \frac{1}{N_t}\right) \mathrm{S}_{tb}^2 + \left(\frac{1}{m} - \frac{1}{M}\right) \frac{\bar{\mathrm{S}}_{tw}^2}{n_t} \qquad (158)$$

whence

$$\text{Est. } \bar{y}_{t..}^2 = \bar{y}_{ts}^2 - \left(\frac{1}{n_t} - \frac{1}{N_t}\right) \hat{\mathrm{S}}_{tb}^2 - \left(\frac{1}{m} - \frac{1}{M}\right) \frac{\hat{\hat{\mathrm{S}}}_{tw}^2}{n_t} \qquad (159)$$

or

$$\text{Est. } \sum_{t=1}^{k} N_t \bar{y}_{t..}^2 = \sum_{t=1}^{k} N_t \bar{y}_{ts}^2 - \sum_{t=1}^{k} N_t \left(\frac{1}{n_t} - \frac{1}{N_t}\right) \hat{\mathrm{S}}_{tb}^2$$

$$- \left(\frac{1}{m} - \frac{1}{M}\right) \sum_{t=1}^{k} \frac{N_t}{n_t} \hat{\hat{\mathrm{S}}}_{tw}^2 \qquad (160)$$

Also

$$V(\bar{y}_w)_{\mathrm{S}} = E\left(\sum_{t=1}^{k} p_t \bar{y}_{ts}\right)^2 - \bar{y}_{..}^2$$

so that

$$\text{Est. } (N\bar{y}_{..}^2) = N\left(\sum_{t=1}^{k} p_t \bar{y}_{ts}\right)^2 - N \cdot \text{Est. } V(\bar{y}_w)_{\mathrm{S}}$$

whence substituting from (156), we have

$$\text{Est. } (N\bar{y}_{..}^2) = N\left(\sum_{t=1}^{k} p_t \bar{y}_{ts}\right)^2 - \sum_{t=1}^{k} N p_t^2 \left\{\left(\frac{1}{n_t} - \frac{1}{N_t}\right) \hat{\mathrm{S}}_{tb}^2\right.$$

$$\left. + \frac{1}{n_t}\left(\frac{1}{m} - \frac{1}{M}\right) \hat{\hat{\mathrm{S}}}_{tw}^2\right\} \qquad (161)$$

Subtracting (161) from (160), we get

$$\text{Est.}\left\{\sum_{t=1}^{k} N_t \bar{y}_{t..}^2 - N\bar{y}_{..}^2\right\} = \sum_{t=1}^{k} N_t\,(\bar{y}_{ts}^2 - \bar{y}_w^2)$$

$$- \left(\frac{1}{m} - \frac{1}{M}\right)\left\{\sum_{t=1}^{k}\left(\frac{N_t}{n_t} - \frac{Np_t^2}{n_t}\right)\hat{\hat{\mathbf{S}}}_{tw}^2\right\}$$

$$- \sum_{t=1}^{k} N_t\left(\frac{1}{n_t} - \frac{1}{N_t}\right)\left(1 - \frac{Np_t^2}{N_t}\right)\hat{\mathbf{S}}_{tb}^2 \tag{162}$$

whence on substituting in (157), we obtain

$$\hat{\mathbf{S}}_b^2 = \frac{1}{N-1}\sum_{t=1}^{k} N_t\,(\bar{y}_{ts}^2 - \bar{y}_w^2)$$

$$+ \frac{1}{N-1}\sum_{t=1}^{k}\left\{N_t - \frac{N_t}{n_t} + \frac{Np_t^2}{n_t} - \frac{Np_t^2}{N_t}\right\}\hat{\mathbf{S}}_{tb}^2$$

$$- \frac{1}{N-1}\left(\frac{1}{m} - \frac{1}{M}\right)\left\{\sum_{t=1}^{k}\frac{N_t}{n_t}\left(1 - \frac{Np_t^2}{N_t}\right)\hat{\hat{\mathbf{S}}}_{tw}^2\right\} \tag{163}$$

If the variation between second-stage units within first-stage units can be assumed to be of the same order from stratum to stratum, we can substitute for $\hat{\hat{\mathbf{S}}}_{tw}^2$ its pooled estimate over all the strata, given by

$$\bar{s}_w^2 = \frac{1}{n\,(m-1)}\sum_{t=1}^{k}\sum_{i}^{n_t}\sum_{j}^{m}(y_{tij} - \bar{y}_{tim})^2 \tag{164}$$

The difference between (153) and (156) when $\mathbf{S}_b^2$ is estimated from (163) and $\mathbf{S}_{tb}^2$ from (150) represents the reduction in variance due to stratification (Sukhatme, 1950).

The difference assumes a simple form when $p_t = N_t/N$. On substituting from (163) in (153), we get

$$\begin{aligned}\text{Est. } V(\bar{y}_{nm})_{US} = \left(\frac{1}{n} - \frac{1}{N}\right) \Bigg[\sum_{t=1}^{k} p_t \hat{S}_{tb}^2 & \\ &+ \frac{N}{N-1} \left\{ \sum_{t=1}^{k} p_t (\bar{y}_{ts} - \bar{y}_w)^2 \right. \\ &\left. - \sum_{t=1}^{k} p_t (1 - p_t) \frac{\hat{S}_{tb}^2}{n_t} \right\} \Bigg] \\ &+ \left(\frac{1}{m} - \frac{1}{M}\right) \frac{1}{n} \Bigg[1 - \frac{N-n}{N-1} \\ &\times \sum_{t=1}^{k} \frac{p_t}{n_t} (1 - p_t) \Bigg] \bar{s}_w^2 \end{aligned} \tag{165}$$

Also, from (156), we have

$$\text{Est. } V(\bar{y}_w)_S = \sum_{t=1}^{k} p_t^2 \left\{ \left(\frac{1}{n_t} - \frac{1}{N_t}\right) \hat{S}_{tb}^2 + \frac{1}{n_t} \left(\frac{1}{m} - \frac{1}{M}\right) \bar{s}_w^2 \right\} \tag{166}$$

Hence

$$\begin{aligned}\text{Est. } \{V_{US} - V_S\} = \frac{N-n}{n(N-1)} \sum_{t=1}^{k} p_t (\bar{y}_{ts} - \bar{y}_w)^2 & \\ &+ \sum_{t=1}^{k} \left\{ \frac{N-n}{Nn} \left(p_t - \frac{N}{N-1} \cdot \frac{p_t(1-p_t)}{n_t} \right) \right. \\ &\left. - \frac{p_t^2}{n_t} + \frac{p_t}{N} \right\} \hat{S}_{tb}^2 \\ &+ \left(\frac{1}{m} - \frac{1}{M}\right) \sum_{t=1}^{k} \left\{ \frac{p_t}{n} - \frac{N-n}{n(N-1)} \right. \\ &\left. \times \frac{p_t(1-p_t)}{n_t} - \frac{p_t^2}{n_t} \right\} \bar{s}_w^2 \end{aligned} \tag{167}$$

Substituting for $\hat{S}_{tb}^2$ from (150) the value

$$s_{tb}^2 - \left(\frac{1}{m} - \frac{1}{M}\right) \bar{s}_w^2$$

we reach the simple expression

$$\text{Est. } \{V_{US} - V_S\} = \frac{N-n}{n\,(N-1)} \sum_{t=1}^{k} p_t\,(\bar{y}_{ts} - \bar{y}_w)^2$$

$$+ \sum_{t=1}^{k} \left\{\frac{p_t}{n} - \frac{p_t^2}{n_t} - \frac{N-n}{N-1} \cdot \frac{1}{n} \cdot \frac{p_t\,(1-p_t)}{n_t}\right\} s_{tb}^2 \tag{168}$$

which is seen to be identical with equation (73) of Chapter III.

REFERENCES

1. Sukhatme, P. V.(1950) .. "Efficiency of Sub-Sampling Designs in Yield Surveys," *Jour. Ind. Soc. Agr. Statist.*, **2**, 212–28.

2. Hansen, M. H. and Hurwitz, W. N. (1943) "On the Theory of Sampling from Finite Populations," *Ann. Math. Statist.*, **14**, 333–62.

3. Sukhatme, P. V. and Panse, V. G. (1951) "Crop Surveys in India—II," *Jour. Ind. Soc. Agr. Statist.*, **3**, 97–168.

4. Yates, F. (1949) .. *Sampling Methods for Censuses and Surveys*, Charles Griffin & Co., Ltd., London.

CHAPTER VIII

SUB-SAMPLING (Continued)

8.1 Introduction

In the preceding chapter we have developed the sampling theory appropriate for sub-sampling systems involving the use of equal probabilities of selection at each stage of sampling. When the first-stage units are large and vary considerably in their sizes, this system of sub-sampling is not usually efficient. This is even more so in cases where practical considerations demand that the survey should be confined to only a small number of first-stage units within each stratum with equal number of second-stage units from each first-stage unit, although the amount of sub-sampling from the selected first-stage units would be necessarily unequal under optimum allocation. A system of sub-sampling involving the use of varying probabilities has been used with considerable gains in efficiency in such cases. In particular, a sub-sampling design in which only one first-stage unit is selected from each stratum, with probability proportional to the measure of the size of the unit, and a fixed number of second-stage units is selected with equal probabilities from each of the selected first-stage units, has been found to bring about marked improvements in precision, compared with sub-sampling systems involving the use of equal probabilities. The developments are due to Hansen and Hurwitz (1943, 1949). In this chapter we shall give the theory of sub-sampling systems involving the use of varying probabilities.

8.2 Estimate of the Population Mean and its Variance

We shall assume that the first-stage units are selected with replacement. Further, we shall suppose that whenever a specified first-stage unit of the population, say the i-th, is included in the sample, a sub-sample of m_i second-stage units will be drawn therefrom without replacement, but only after the replacement of any sub-sample which may have been drawn previously. In other words, if the i-th unit happens to be selected, say γ times, in a sample of n first-stage units, γ sub-samples of m_i units each will be drawn therefrom independently of each other, each sub-sample of m_i being drawn without replacement.

Let P_i denote the selection probability assigned to the i-th first-stage unit of the population ($i = 1, 2, \ldots, N$) and $\sum_{i=1}^{N} P_i = 1$. Further, let

$$z_{ij} = \frac{M_i}{M_0} \cdot \frac{1}{P_i} y_{ij} \tag{1}$$

Consider the estimate

$$\begin{aligned} \bar{z}_s &= \bar{z}_{n(m_i)} \\ &= \frac{1}{n} \sum^{n} \bar{z}_{i(m_i)} \end{aligned} \tag{2}$$

where the summation is taken over all the n units in the sample. Then it is easily shown that $\bar{z}_s$ is an unbiased estimate of $\bar{y}_{..}$. For, we have

$$\begin{aligned} E(\bar{z}_s) &= E\left\{\frac{1}{n} \sum^{n} \bar{z}_{i(m_i)}\right\} \\ &= E\left\{\frac{1}{n} \sum^{n} E(\bar{z}_{i(m_i)} \mid i)\right\} \\ &= E\left\{\frac{1}{n} \sum^{n} \bar{z}_{i.}\right\} \\ &= \frac{1}{n} \sum^{n} E(\bar{z}_{i.}) \\ &= \bar{z}_{..} \end{aligned} \tag{3}$$

where

$$\bar{z}_{..} = \sum_{i=1}^{N} P_i \bar{z}_{i.} \tag{4}$$

$$\begin{aligned} &= \sum_{i=1}^{N} \frac{M_i}{M_0} \bar{y}_{i.} \\ &= \bar{y}_{..} \end{aligned} \tag{5}$$

To evaluate the sampling variance of $\bar{z}_s$, we write

$$\begin{aligned} V(\bar{z}_s) &= E\{\bar{z}_s - E(\bar{z}_s)\}^2 \\ &= E(\bar{z}_s - \bar{z}_{..})^2 \\ &= E(\bar{z}_{n(m_i)} - \bar{z}_{n.} + \bar{z}_{n.} - \bar{z}_{..})^2 \\ &= E(\bar{z}_{n(m_i)} - \bar{z}_{n.})^2 + E(\bar{z}_{n.} - \bar{z}_{..})^2 \\ &\qquad + 2E(\bar{z}_{n(m_i)} - \bar{z}_{n.})(\bar{z}_{n.} - \bar{z}_{..}) \end{aligned} \tag{6}$$

Taking the first term in (6), we have

$$\begin{aligned} E(\bar{z}_{n(m_i)} - \bar{z}_{n.})^2 &= E\left\{\frac{1}{n}\sum^{n}(\bar{z}_{i(m_i)} - \bar{z}_{i.})\right\}^2 \\ &= \frac{1}{n^2}E\left\{\sum^{n}(\bar{z}_{i(m_i)} - \bar{z}_{i.})^2\right. \\ &\qquad \left. + \sum_{i \neq i'}^{n}(\bar{z}_{i(m_i)} - \bar{z}_{i.})(\bar{z}_{i'(m_{i'})} - \bar{z}_{i'.})\right\} \end{aligned}$$

Since the first-stage units are sub-sampled independently of each other, we may write

$$\begin{aligned} E(\bar{z}_{n(m_i)} - \bar{z}_{n.})^2 &= \frac{1}{n^2}\left\{\sum^{n}E(\bar{z}_{i(m_i)} - \bar{z}_{i.})^2\right. \\ &\qquad \left. + \sum_{i \neq i'}^{n}E(\bar{z}_{i(m_i)} - \bar{z}_{i.}) \cdot E(\bar{z}_{i'(m_{i'})} - \bar{z}_{i'.})\right\} \\ &= \frac{1}{n}\left\{\sum_{i=1}^{N}P_i\left(\frac{1}{m_i} - \frac{1}{M_i}\right)S_{iz}^2\right\} \end{aligned} \tag{7}$$

where

$$\begin{aligned} S_{iz}^2 &= \frac{1}{M_i - 1}\sum_{j=1}^{M_i}(z_{ij} - \bar{z}_{i.})^2 \\ &= \frac{M_i^2}{M_0^2 P_i^2} \cdot \frac{1}{M_i - 1}\sum_{j=1}^{M_i}(y_{ij} - \bar{y}_{i.})^2 \\ &= \frac{M_i^2}{M_0^2 P_i^2} S_i^2 \end{aligned} \tag{8}$$

The value of the second term in (6) is given by (77) of Chapter VI. We have

$$E\,(\bar{z}_{n.}-\bar{z}_{..})^2 = \frac{\sigma_{bz}{}^2}{n} \tag{9}$$

where

$$\begin{aligned}\sigma_{bz}{}^2 &= \sum_{i=1}^{N} P_i\,(\bar{z}_{i.}-\bar{z}_{..})^2\\ &= \sum_{i=1}^{N} P_i\left(\frac{M_i\bar{y}_{i.}}{M_0P_i}-\bar{y}_{..}\right)^2\\ &= \sum_{i=1}^{N} \frac{M_i{}^2\bar{y}_{i.}{}^2}{M_0{}^2P_i} - \bar{y}_{..}{}^2 \end{aligned} \tag{10}$$

The expected value of the third term in (6) is obviously zero. We therefore obtain

$$V\,(\bar{z}_s) = \frac{\sigma_{bz}{}^2}{n} + \frac{1}{n}\sum_{i=1}^{N} P_i\left(\frac{1}{m_i}-\frac{1}{M_i}\right)\mathbf{S}_{iz}{}^2 \tag{11}$$

which can also be alternatively written as

$$V\,(\bar{z}_s) = \frac{1}{n}\left\{\sum_{i=1}^{N}\frac{M_i{}^2\bar{y}_{i.}{}^2}{M_0{}^2P_i} - \bar{y}_{..}{}^2\right\} + \frac{1}{n}\sum_{i=1}^{N}\frac{M_i{}^2}{M_0{}^2\,P_i}\left(\frac{1}{m_i}-\frac{1}{M_i}\right)\mathbf{S}_i{}^2 \tag{12}$$

When the selection probabilities are such that

$$P_i = \frac{M_i}{M_0} \qquad (i = 1, 2, \ldots, N)$$

the expression for $\sigma_{bz}{}^2$ is simplified, being given by

$$\sigma_{bz}{}^2 = \sum_{i=1}^{N}\frac{M_i}{M_0}\,(\bar{y}_{i.}-\bar{y}_{..})^2 = \sigma_{by}{}^2 \tag{13}$$

whence

$$V(\bar{z}_s) = \frac{1}{n} \sum_{i=1}^{N} \frac{M_i}{M_0} (\bar{y}_{i.} - \bar{y}_{..})^2 + \frac{1}{n} \sum_{i=1}^{N} \frac{M_i}{M_0} \left(\frac{1}{m_i} - \frac{1}{M_i}\right) \mathrm{S}_i^2 \tag{14}$$

Whenever the mean square and the variance relate to the variate z, we have so indicated by adding z as a subscript to S and σ. In all other cases, S and σ should be taken to relate to the variate y only.

8.3 Estimation of the Variance from the Sample

Consider the mean square of $\bar{z}_{i(m_i)}$ obtained from the sample

$$s_{bz}^2 = \frac{1}{n-1} \sum^{n} (\bar{\bar{z}}_{i(m_i)} - \bar{\bar{z}}_{n(m_i)})^2 \tag{15}$$

Expanding and taking expectations, we get

$$E(s_{bz}^2) = \frac{1}{n-1} E \left\{ \sum^{n} \bar{\bar{z}}^2{}_{i(m_i)} - n\bar{\bar{z}}^2{}_{n(m_i)} \right\}$$

$$= \frac{1}{n-1} \left\{ \sum^{n} E(\bar{\bar{z}}^2{}_{i(m_i)}) - nE(\bar{\bar{z}}^2{}_{n(m_i)}) \right\}. \tag{16}$$

Since the first-stage units are selected with replacement, we may consider our sample of n to be the result of n independent samples of one each, and write the right-hand side in (16) as

$$= \frac{1}{n-1} \left[n \{ V(\bar{\bar{z}}_{i(m_i)}) + \bar{\bar{z}}_{..}^2 \} - n \{ V(\bar{\bar{z}}_{n(m_i)}) + \bar{\bar{z}}_{..}^2 \} \right] \tag{17}$$

where $V(\bar{z}_{i(m_i)})$ is the variance of the mean based on a sample of one first-stage unit.

Substituting from (11) in (17), we get

$$E(s_{bz}^2) = \sigma_{bz}^2 + \sum_{i=1}^{N} P_i \left(\frac{1}{m_i} - \frac{1}{M_i}\right) \mathrm{S}_{iz}^2 \tag{18}$$

It follows that

$$\text{Est. } V(\bar{z}_s) = \frac{s_{bz}^2}{n} \tag{19}$$

8.4 Allocation of Sample

So far we have placed no restriction on the size of the sub-sample to be drawn from a selected first-stage unit. It may be related to the selected unit or independent of it. The guiding principle in determining the optimum values of n and m_i is clearly to maximize the precision for a given cost, or to minimize the cost for a desired precision.

In a large number of surveys, the cost will be determined by the number of different first-stage units and the total number of second-stage units in the sample. We shall suppose here that the cost of the survey is made up of two components, as follows:

$$C = c_1 n' + c_2 \sum^{n} m_i \tag{20}$$

where

n' denotes the different first-stage units included in the sample,

$\sum^{n} m_i$ the total number of second-stage units in the sample,

c_1 the cost per first-stage unit on travel and setting up of an office,

and

c_2 the cost per second-stage unit of collecting the required information.

Clearly, C will vary from sample to sample of n. In actual practice, one must be able to predict the cost in advance of designing the sample in order to be able to compute the optima. We shall, therefore, consider the average instead of the actual cost of surveying a sample. Now to obtain the average value of the first term in (20), we have

$$c_1 E(n') = c_1 \sum_{i=1}^{N} 1 \cdot \{\text{Probability that the } i\text{-th unit is included at least once in a sample of } n\}$$

$$= c_1 \sum_{i=1}^{N} 1 \cdot \{1 - \text{Probability that the } i\text{-th unit is not included in any of the } n \text{ draws}\}$$

$$= c_1 \sum_{i=1}^{N} \{1 - (1 - P_i)^n\} \tag{21}$$

The average value of the second component in (20) is given by

$$c_2 E\left(\sum^{n} m_i\right) = c_2 n \sum_{i=1}^{N} P_i m_i \tag{22}$$

Hence the average total cost of the survey will be represented by

$$C = c_1 \sum_{i=1}^{N} \{1 - (1 - P_i)^n\} + c_2 n \sum_{i=1}^{N} P_i m_i \tag{23}$$

This cost function, however, offers a slight disadvantage in that it is not simple to deal with. We shall, therefore, suppose that N is reasonably large and none of the P_i's too large, so that the average cost may be approximated by the simple function

$$C = c_1 n + c_2 n \sum_{i=1}^{N} P_i m_i \tag{24}$$

To determine the optima, the simplest method would be to consider the product of (11) with (24), and write

$$V(\bar{z}_s) \cdot C = \left\{\sigma_{bz}^2 + \sum_{i=1}^{N} P_i \left(\frac{1}{m_i} - \frac{1}{M_i}\right) \mathbf{S}_{iz}^2\right\} \times \left\{c_1 + c_2 \sum_{i=1}^{N} P_i m_i\right\} \tag{25}$$

Clearly, the minimum value of (25) will produce optima when either C or V is fixed in advance and V or C is minimized.

Let

$$\Delta = \sigma_{bz}^2 - \sum_{i=1}^{N} P_i \frac{\mathbf{S}_{iz}^2}{M_i} \tag{26}$$

Equation (25) can then be rewritten and expanded as

$$V(\bar{z}_s) \cdot C = \left\{\varDelta + \sum_{i=1}^{N} P_i \frac{\mathrm{S}_{iz}^{\,2}}{m_i}\right\} \left\{c_1 + c_2 \sum_{i=1}^{N} P_i m_i\right\}$$

$$= c_1 \varDelta + c_2 \sum_{i=1}^{N} P_i^{\,2} \mathrm{S}_{iz}^{\,2} + \sum_{i=1}^{N} P_i \left\{c_1 \frac{\mathrm{S}_{iz}^{\,2}}{m_i} + c_2 \varDelta m_i\right\}$$

$$+ c_2 \sum_{i>i'=1}^{N} P_i P_{i'} \left\{\frac{\mathrm{S}_{iz}^{2}}{m_i} m_{i'} + \frac{\mathrm{S}_{i'z}^{\,2}}{m_{i'}} m_i\right\} \qquad (27)$$

Assuming $\varDelta$ to be positive, (27) can be put in the form

$$V(\bar{z}_s) \cdot C = c_1 \varDelta + c_2 \sum_{i=1}^{N} P_i^{2} \mathrm{S}_{iz}^{\,2} + 2 \sum_{i=1}^{N} P_i \sqrt{c_1 c_2 \varDelta \, \mathrm{S}_{iz}^{\,2}}$$

$$+ 2c_2 \sum_{i>i'=1}^{N} P_i P_{i'} \mathrm{S}_{iz} \mathrm{S}_{i'z}$$

$$+ \sum_{i=1}^{N} P_i \left\{\sqrt{\frac{c_1 \mathrm{S}_{iz}^{\,2}}{m_i}} - \sqrt{c_2 \varDelta m_i}\right\}^2$$

$$+ c_2 \sum_{i>i'=1}^{N} \left\{\sqrt{\frac{m_{i'}}{m_i} \; P_i P_{i'} \mathrm{S}_{iz}^{\,2}}\right.$$

$$\left. - \sqrt{\frac{m_i}{m_{i'}} \; P_i P_{i'} \mathrm{S}_{i'z}^{\,2}}\right\}^2 \qquad (28)$$

This is minimum when each of the two square terms in (28) is zero, giving us for the optimum of m_i the value

$$m_i = \sqrt{\frac{c_1}{c_2 \varDelta}} \, \mathrm{S}_{iz}$$

$$= \sqrt{\frac{c_1}{c_2 \varDelta}} \cdot \frac{M_i}{M_0 P_i} \mathrm{S}_i \qquad (i = 1, 2, \ldots, N) \qquad (29)$$

In other words, m_i should be so determined that

$$\frac{P_i m_i}{M_i \mathrm{S}_i} = \text{constant}$$

$$= \frac{1}{M_0}\sqrt{\frac{c_1}{c_2 \Delta}} \qquad (i = 1, 2, \ldots, N) \qquad (30)$$

However, S_i will ordinarily not be known. It will also vary from character to character. Practical considerations require that m_i should be independent of S_i even if it means somewhat departing from the optimum. In practice, S_i will be generally found to increase with M_i, though seldom as fast as M_i. We shall assume here that S_i is a constant equal to S_w, say, in which case m_i would be so determined that

$$P_i \frac{m_i}{M_i} = k$$

$$= \text{a constant} \qquad (i = 1, 2, \ldots, N) \qquad (31)$$

Knowing k, the value of n is obtained from (11) or (24), depending upon whether the cost of the survey is minimized for fixed V_0, or the variance is minimized for fixed C_0. In the former case,

$$\hat{n} = \frac{\sigma_{bz}^2 + \sum_{i=1}^{N} \frac{P_i}{M_i}\left(\frac{P_i}{k} - 1\right)\mathrm{S}_{iz}^2}{V_0} \qquad (32)$$

and in the latter,

$$\hat{n} = \frac{C_0}{c_1 + c_2 k M_0}. \qquad (33)$$

We remark that when P_i is proportional to M_i, the optimum value of m_i is a constant, irrespective of the first-stage unit included in the sample.

8.5* Determination of Optimum Probabilities

In determining the optimum allocation of the sample in the previous section we assumed that the selection probabilities P_i were given. These probabilities can be any arbitrary positive proper fractions, subject to the condition that their sum is 1;

alternatively, they can be related in a known way to the characteristics of the units to be selected.

The optimum values of selection probabilities are given by minimizing the variance of $\bar{z}_s$ for given cost. In this section we shall determine them assuming that: (*a*) the sub-sampling rate m_i/M_i for a specified first-stage unit i will be such that equation (31) is satisfied, and (*b*) the cost function is independent of P_i's.

Following the Lagrange procedure, we consider the function ϕ given by

$$\phi = V(\bar{z}_s) + \lambda\left(\sum_{i=1}^{N} P_i - 1\right) \tag{34}$$

where λ is a constant multiplier. Now the value of $V(\bar{z}_s)$ in terms of P_i's is obtained from (12) after substituting for m_i from (31), and is given by

$$V(\bar{z}_s) = \frac{1}{n}\left\{\sum_{i=1}^{N} \frac{M_i^2 \bar{y}_{i.}^2}{M_0^2 P_i} - \bar{y}_{..}^2\right\} + \frac{1}{nkM_0^2}\sum_{i=1}^{N} M_i \mathrm{S}_i^2$$

$$- \frac{1}{nM_0^2}\sum_{i=1}^{N} \frac{M_i}{P_i}\mathrm{S}_i^2 \tag{35}$$

Substituting from (35) in (34), we write

$$\phi = \frac{1}{n}\left\{\sum_{i=1}^{N} \frac{M_i^2 \bar{y}_{i.}^2}{M_0^2 P_i} - \bar{y}_{..}^2\right\} + \frac{1}{nkM_0^2}\sum_{i=1}^{N} M_i \mathrm{S}_i^2$$

$$- \frac{1}{nM_0^2}\sum_{i=1}^{N} \frac{M_i}{P_i}\mathrm{S}_i^2 + \lambda\left(\sum_{i=1}^{N} P_i - 1\right)$$

Differentiating ϕ with respect to P_i and equating to zero gives

$$\frac{\partial\phi}{\partial P_i} = -\frac{1}{nM_0^2}\left\{\frac{M_i^2 \bar{y}_{i.}^2}{P_i^2} - \frac{M_i \mathrm{S}_i^2}{P_i^2}\right\} + \lambda = 0$$

Hence

$$P_i = \frac{M_i \bar{y}_{i.}\sqrt{1 - \dfrac{\mathrm{S}_i^2}{M_i \bar{y}_{i.}^2}}}{\sqrt{\lambda n M_0^2}} \qquad (i = 1, 2, \ldots, N)$$

But $\sum\limits_{i=1}^{N} P_i = 1$. We, therefore, have

$$P_i = \frac{M_i\bar{y}_{i.}\sqrt{1 - \dfrac{S_i^2}{M_i\bar{y}_{i.}^2}}}{\sum\limits_{i=1}^{N} M_i\bar{y}_{i.}\sqrt{1 - \dfrac{S_i^2}{M_i\bar{y}_{i.}^2}}} \qquad (i = 1, 2, \ldots, N) \tag{36}$$

It will be noticed that the optimum value of P_i depends on two factors: (i) the total of the character for the i-th first-stage unit, and (ii) the coefficient of variation of the first-stage unit mean. The latter will usually be a very small fraction so that P_i will be primarily determined by the first-stage unit total $M_i\bar{y}_{i.}$. In practice, however, $M_i\bar{y}_{i.}$ will not be known although quantities correlated with them, as for example those determined from the previous census, may be available. Failing to have these, it would appear that the choice of P_i proportional to M_i would give about the optimum probabilities.

It should be pointed out that the above solution for optimum probabilities will hold even when the cost function is represented by (24); for, under the assumption that $P_i m_i/M_i = k$, (24) is seen to be independent of P_i's. However, in a number of situations the survey cost may not be independent of P_i's. Consider, for example, a situation where a list of first-stage units is available but that for the second-stage units within first-stage units is not. Listing of the second-stage units within the selected first-stage units is, therefore, an essential part of the survey work. The situation is of common occurrence in agricultural surveys in under-developed countries. Thus, lists of villages are readily available in most countries and the identification of boundaries of selected villages also does not present any difficulty. However, lists of second-stage units like fields or families are not available and have to be prepared for selecting a subsample. To the cost of survey represented by (20), we, therefore, have to add a component for expenditure for listing. This will generally vary with the size of the first-stage units. We, therefore, have a cost function given by

$$C = c_1 n + c_2 \sum^{n} m_i + c_3 \sum^{n} M_i \tag{37}$$

where c_3 represents the cost of listing a second-stage unit in a selected first-stage unit. The average value of C in repeated samples will be given by

$$c_1 n + c_2 n \sum_{i=1}^{N} P_i m_i + c_3 n \sum_{i=1}^{N} P_i M_i \tag{38}$$

and is seen to reduce to

$$c_1 n + c_2 nkM_0 + c_3 n \sum_{i=1}^{N} P_i M_i \tag{39}$$

for $P_i m_i = kM_i$. The cost function (39) will now be seen to depend upon P_i's. However, if M_i's are unknown, M_0 will also not be known and the estimate $\bar{z}_s$ can no longer be used. Several alternative estimates can be formed. We shall consider one such estimate in this chapter, namely, the ratio estimate, and thereafter resume discussion of the problem considered in this section.

8.6* Ratio Estimate

Let

$$\bar{y}_R = \frac{\bar{z}_s}{\bar{v}_s}\bar{v}_{..} = R_s \bar{v}_{..} \tag{40}$$

denote the ratio estimate of the population mean, where

$$\left.\begin{aligned} z_{ij} &= \frac{M_i}{M_0 P_i} y_{ij} \qquad & \bar{z}_s &= \frac{1}{n}\sum^{n} \frac{M_i}{M_0 P_i} \bar{y}_{i(m_i)} \\ v_{ij} &= \frac{M_i}{M_0 P_i} x_{ij} \qquad & \bar{v}_s &= \frac{1}{n}\sum^{n} \frac{M_i}{M_0 P_i} \bar{x}_{i(m_i)} \end{aligned}\right\} \tag{41}$$

x standing for a supplementary variate observed for all units in the sample.

Then from the results of Chapter IV we may, for n sufficiently large, ignore the bias terms in the expected value of $\bar{y}_R$ and regard it as an unbiased estimate of the population mean for all practical purposes.

To obtain the variance of the estimate $\bar{y}_R$, we shall start with the expression (28) in Chapter IV. Since $E(\bar{z}_S) = \bar{y}_{..}$ and $E(\bar{v}_S) = \bar{x}_{..}$, we write to a first approximation

$$V(\bar{y}_R) = \bar{y}_{..}^2 \left\{ \frac{V(\bar{z}_s)}{\bar{y}_{..}^2} + \frac{V(\bar{v}_s)}{\bar{x}_{..}^2} - \frac{2\,\mathrm{Cov}\,(\bar{z}_s, \bar{v}_s)}{\bar{y}_{..}\,\bar{x}_{..}} \right\}. \tag{42}$$

Now, from (12), we have

$$V(\bar{z}_s) = \frac{1}{n} \left(\sum_{i=1}^{N} \frac{M_i^2 \bar{y}_{i.}^2}{M_0^2 P_i} - \bar{y}_{..}^2 \right) + \frac{1}{nN^2} \sum_{i=1}^{N} \frac{M_i^2}{\bar{M}^2 P_i} \left(\frac{1}{m_i} - \frac{1}{M_i} \right) \mathrm{S}_{iy}^2 \tag{43}$$

and by analogy

$$V(\bar{v}_s) = \frac{1}{n} \left(\sum_{i=1}^{N} \frac{M_i^2 \bar{x}_{i.}^2}{M_0^2 P_i} - \bar{x}_{..}^2 \right) + \frac{1}{nN^2} \sum_{i=1}^{N} \frac{M_i^2}{\bar{M}^2 P_i} \left(\frac{1}{m_i} - \frac{1}{M_i} \right) \mathrm{S}_{ix}^2 \tag{44}$$

Further,

$$\begin{aligned} \mathrm{Cov}\,(\bar{z}_s, \bar{v}_s) &= E\{(\bar{z}_s - \bar{z}_{..})(\bar{v}_s - \bar{v}_{..})\} \\ &= E\{(\bar{z}_{n(m_i)} - \bar{z}_{n.} + \bar{z}_{n.} - \bar{z}_{..}) \\ &\qquad \times (\bar{v}_{n(m_i)} - \bar{v}_{n.} + \bar{v}_{n.} - \bar{v}_{..})\} \\ &= E\{(\bar{z}_{n(m_i)} - \bar{z}_{n.})(\bar{v}_{n(m_i)} - \bar{v}_{n.}) \\ &\qquad + (\bar{z}_{n.} - \bar{z}_{..})(\bar{v}_{n.} - \bar{v}_{..})\} \end{aligned} \tag{45}$$

since the expectations of the other two product terms are zero.

Now taking the first term in (45), we have

$$\begin{aligned} &E\{(\bar{z}_{n(m_i)} - \bar{z}_{n.})(\bar{v}_{n(m_i)} - \bar{v}_{n.})\} \\ &= E\left[\left\{ \frac{1}{n} \sum^{n} \frac{M_i}{M_0 P_i} (\bar{y}_{i(m_i)} - \bar{y}_{i.}) \right\} \right. \\ &\qquad \left. \times \left\{ \frac{1}{n} \sum^{n} \frac{M_i}{M_0 P_i} (\bar{x}_{i(m_i)} - \bar{x}_{i.}) \right\} \right] \end{aligned}$$

$$= \frac{1}{n^2} E\left[\sum^{n} \frac{M_i^2}{M_0^2 P_i^2} (\bar{y}_{i(m_i)} - \bar{y}_{i.}) (\bar{x}_{i(m_i)} - \bar{x}_{i.}) + \sum_{i \neq i'}^{n} \frac{M_i M_{i'}}{M_0^2 P_i P_{i'}} (\bar{y}_{i(m_i)} - \bar{y}_{i.}) (\bar{x}_{i'(m_{i'})} - \bar{x}_{i'.})\right]$$

$$= \frac{1}{n^2 M_0^2} E\left[\sum^{n} \frac{M_i^2}{P_i^2} E\{(\bar{y}_{i(m_i)} - \bar{y}_{i.}) (\bar{x}_{i(m_i)} - \bar{x}_{i.}) \mid i\}\right]$$

$$= \frac{1}{n^2 M_0^2} E\left[\sum^{n} \frac{M_i^2}{P_i^2} \left(\frac{1}{m_i} - \frac{1}{M_i}\right) \mathrm{S}_{iyx}\right]$$

$$= \frac{1}{n M_0^2} \sum_{i=1}^{N} \frac{M_i^2}{P_i} \left(\frac{1}{m_i} - \frac{1}{M_i}\right) \mathrm{S}_{iyx} \qquad (46)$$

The second term in (45) is clearly given by

$$E(\bar{z}_{n.} - \bar{z}_{..}) (\bar{v}_{n.} - \bar{v}_{..})$$

$$= \frac{1}{n} \mathrm{S}_{bzv}$$

$$= \frac{1}{n} \sum_{i=1}^{N} P_i \left(\frac{M_i}{M_0 P_i} \bar{y}_{i.} - \bar{y}_{..}\right) \left(\frac{M_i}{M_0 P_i} \bar{x}_{i.} - \bar{x}_{..}\right)$$

$$= \frac{1}{n} \left\{\sum_{i=1}^{N} \frac{M_i^2}{M_0^2 P_i} \bar{y}_{i.} \bar{x}_{i.} - \bar{y}_{..} \bar{x}_{..}\right\} \qquad (47)$$

Substituting from (46) and (47) in (45), we have

$$\mathrm{Cov}\ (\bar{z}_s, \bar{v}_s) = \frac{1}{n} \left\{\frac{1}{M_0^2} \sum_{i=1}^{N} \frac{M_i^2}{P_i} \bar{y}_{i.} \bar{x}_{i.} - \bar{y}_{..} \bar{x}_{..}\right\} + \frac{1}{n M_0^2} \sum_{i=1}^{N} \frac{M_i^2}{P_i} \left(\frac{1}{m_i} - \frac{1}{M_i}\right) \mathrm{S}_{iyx} \qquad (48)$$

Hence, from (42), we obtain

$$V(\bar{y}_R) \cong \frac{1}{n} \left\{ \sum_{i=1}^{N} \frac{M_i^2}{M_0^2 P_i} \left(\bar{y}_{i.} - \frac{\bar{y}_{..}}{\bar{x}_{..}} \bar{x}_{i.} \right)^2 \right\}$$

$$+ \frac{1}{nM_0^2} \sum_{i=1}^{N} \frac{M_i^2}{P_i} \left(\frac{1}{m_i} - \frac{1}{M_i} \right) D_i^2 \qquad (49)$$

where

$$D_i^2 = \mathbf{S}_{iy}^2 + R^2 \mathbf{S}_{ix}^2 - 2R\mathbf{S}_{iyx} \qquad (50)$$

and

$$R = \frac{\bar{y}_{..}}{\bar{x}_{..}}$$

When $x_{ij} = 1$, the expression for the estimate $\bar{y}_R$ is simplified, being given by

$$\bar{y}_R = \frac{\bar{z}_s}{\bar{u}_s} \qquad (51)$$

where

$$u_i = \frac{M_i}{M_0 P_i} \quad \text{and} \quad \bar{u}_s = \frac{1}{n} \sum^{n} u_i \qquad (52)$$

Also,

$$\mathbf{S}_{ix}^2 = 0$$

and the variance of $\bar{y}_R$ takes the simple form given by

$$V(\bar{y}_R) \cong \frac{1}{nM_0^2} \left\{ \sum_{i=1}^{N} \frac{M_i^2}{P_i} (\bar{y}_{i.} - \bar{y}_{..})^2 \right.$$

$$\left. + \sum_{i=1}^{N} \frac{M_i^2}{P_i} \left(\frac{1}{m_i} - \frac{1}{M_i} \right) \mathbf{S}_{iy}^2 \right\} \qquad (53)$$

An estimate of $V(\bar{y}_R)$ is easily obtained. The reader may verify that, following the steps shown in Section 7.13, the estimate is given by

$$\text{Est. } V(\bar{y}_R) \cong \frac{1}{n} \cdot \frac{1}{n-1} \sum^{n} \frac{M_i^2}{M_0^2 P_i^2} (\bar{y}_{i(m_i)} - R_s \bar{x}_{i(m_i)})^2 \qquad (54)$$

where

$$R_s = \text{Est. } R = \frac{\bar{z}_s}{\bar{v}_s}$$

8.7* Allocation of Sample and Determination of Optimum Probabilities: General Case

We shall now return to the problems discussed in Sections 8.4 and 8.5, namely, to determine the optimum values of m_i's and P_i's when the cost function is represented by (39) and the estimate used is the ratio estimate $\bar{y}_R$. The solution is straightforward since we note that: (*a*) $V(\bar{y}_R)$ has the same form as $V(\bar{z}_S)$ except that instead of σ_{bz}^2, we now have

$$\sum_{i=1}^{N} P_i (\bar{z}_{i.} - R\bar{v}_{i.})^2 \tag{55}$$

where

$$R = \frac{\bar{y}_{..}}{\bar{x}_{..}} \tag{56}$$

and instead of S_{iz}^2 we have $M_i^2 D_i^2 / M_0^2 P_i^2$, and (*b*) the cost function (38) regarded as a function of m_i has the same relationship to m_i as the one represented by (24) has with m_i, except that c_1 is now replaced by c_1', where

$$c_1' = c_1 + c_3 \sum_{i=1}^{N} P_i M_i \tag{57}$$

It follows from (29) that the optimum value of m_i is now determined by

$$m_i = \frac{1}{M_0} \sqrt{\frac{c_1'}{c_2 \Delta'}} \cdot \frac{M_i D_i}{P_i} \quad (i = 1, 2, \ldots, N) \tag{58}$$

where

$$\Delta' = \sum_{i=1}^{N} P_i (\bar{z}_{i.} - R\bar{v}_{i.})^2 - \frac{1}{M_0^2} \sum_{i=1}^{N} \frac{M_i}{P_i} D_i^2 \tag{59}$$

If D_i is constant we reach the same result as (31), namely,

$$\frac{P_i m_i}{M_i} = k$$

To determine the optimum probabilities we form the function ϕ given by

$$\phi = V(\bar{y}_R) + \mu\,(C - C_0) + \lambda\,\left(\sum_{i=1}^{N} P_i - 1\right) \tag{60}$$

where μ and λ are Lagrangian constants. Substituting for $V(\bar{y}_R)$ and C in terms of P_i, k and n, we obtain

$$\begin{aligned}\phi = \frac{1}{nM_0^2}\left\{\sum_{i=1}^{N} \frac{M_i^2}{P_i}(\bar{y}_{i.} - R\bar{x}_{i.})^2 + \frac{1}{k}\sum_{i=1}^{N} M_i D_i^2 \right.\\ \left. - \sum_{i=1}^{N} \frac{M_i}{P_i} D_i^2\right\}\\ + \mu\left\{n\left(c_1 + c_2kM_0 + c_3 \sum_{i=1}^{N} P_iM_i\right) - C_0\right\} + \lambda\left(\sum_{i=1}^{N} P_i - 1\right)\end{aligned} \tag{61}$$

Differentiating ϕ with respect to P_i, k and n and equating to zero gives

$$\begin{aligned}\frac{\partial\phi}{\partial P_i} = -\frac{1}{nM_0^2}\left\{\frac{M_i^2}{P_i^2}(\bar{y}_{i.} - R\bar{x}_{i.})^2 - \frac{M_i}{P_i^2} D_i^2\right\}\\ + \mu nc_3M_i + \lambda = 0 \qquad (i = 1, 2, \ldots, N)\end{aligned} \tag{62}$$

$$\frac{\partial\phi}{\partial k} = -\frac{1}{nM_0^2k^2}\sum_{i=1}^{N} M_iD_i^2 + \mu nc_2M_0 = 0 \tag{63}$$

$$\begin{aligned}\frac{\partial\phi}{\partial n} = -\frac{1}{n^2M_0^2}\left\{\sum_{i=1}^{N} \frac{M_i^2}{P_i}(\bar{y}_{i.} - R\bar{x}_{i.})^2 \right.\\ \left. + \frac{1}{k}\sum_{i=1}^{N} M_iD_i^2 - \sum_{i=1}^{N} \frac{M_i}{P_i} D_i^2\right\}\\ + \mu\left(c_1 + c_2kM_0 + c_3\sum_{i=1}^{N} P_iM_i\right) = 0\end{aligned} \tag{64}$$

The solutions of $N + 2$ equations (62), (63) and (64) subject to the two conditions imposed by the fixed cost and the sum of P_i's being unity, give the optimum values of P_i, n and k. To solve these, we multiply (62) by P_i and sum over all the N units, giving us

$$\frac{1}{nM_0^2}\left\{\sum_{i=1}^{N}\frac{M_i^2}{P_i}(\bar{y}_{i.} - R\bar{x}_{i.})^2 - \sum_{i=1}^{N}\frac{M_i}{P_i}D_i^2\right\}$$

$$= \mu nc_3 \sum_{i=1}^{N} P_iM_i + \lambda \tag{65}$$

From (63), we have

$$\frac{1}{nkM_0^2}\sum_{i=1}^{N} M_iD_i^2 = \mu nkc_2M_0 \tag{66}$$

while (64) gives

$$\frac{1}{nM_0^2}\left\{\sum_{i=1}^{N}\frac{M_i^2}{P_i}(\bar{y}_{i.} - R\bar{x}_{i.})^2 + \frac{1}{k}\sum_{i=1}^{N} M_iD_i^2 - \sum_{i=1}^{N}\frac{M_i}{P_i}D_i^2\right\}$$

$$= \mu\,(c_1n + c_2nkM_0 + c_3n \sum_{i=1}^{N} P_iM_i) \tag{67}$$

Subtracting (65) and (66) from (67), we get

$$\lambda = \mu c_1 n$$

whence substituting for λ in (62), we get

$$P_i \propto M_i \sqrt{\frac{(\bar{y}_{i.} - R\bar{x}_{i.})^2 - \dfrac{D_i^2}{M_i}}{c_1 + c_3M_i}} \tag{68}$$

$$= \frac{M_i \sqrt{\dfrac{(\bar{y}_{i.} - R\bar{x}_{i.})^2 - \dfrac{D_i^2}{M_i}}{c_1 + c_3M_i}}}{\sum\limits_{i=1}^{N} M_i \sqrt{\dfrac{(\bar{y}_{i.} - R\bar{x}_{i.})^2 - \dfrac{D_i^2}{M_i}}{c_1 + c_3M_i}}} \tag{69}$$

It will be noticed that the optimum value of P_i now depends upon (i) the size of the first-stage unit, and (ii) a quantity

$$\delta_i = (\bar{y}_{i.} - R\bar{x}_{i.})^2 - \frac{D_i^2}{M_i} \tag{70}$$

Hansen and Hurwitz (1943) have presented evidence which shows that δ_i tends to decrease as M_i increases although not as fast as M_i. Assuming δ_i to be constant, it will be seen that for $c_3 = 0$, P_i varies as M_i, thus confirming the result reached in Section 8.5. When $c_1 = 0$ and $c_3 > 0$, probability proportional to the square root of M_i will appear to be the optimum.

Solving the other equations the reader may verify that the values of k and n are given by

$$k = \frac{\sqrt{\dfrac{1}{M_0} \sum\limits_{i=1}^{N} M_i D_i^2}}{\sum\limits_{i=1}^{N} \sqrt{\dfrac{M_i^2 \delta_i}{c_1 + c_3 M_i} c_2}} \tag{71}$$

and

$$n = \frac{C_0}{c_1 + c_2 k M_0 + c_3 \sum\limits_{i=1}^{N} P_i M_i} \tag{72}$$

The optima will naturally vary with the cost function and the sampling system, and care is necessary to determine from pilot studies the nature of the cost functions before deciding on the optima to be adopted for the surveys.

8.8 Relative Efficiency of the Two Sub-Sampling Designs

We remarked in the introduction to this chapter that a sub-sampling design in which the selection probabilities are proportional to the size of the first-stage units, and a constant number of second-stage units is drawn from each selected first-stage unit, may bring about a marked improvement in precision compared to the sub-sampling design involving the use of equal selection probabilities. In this section we shall compare the two systems, using $\bar{z}_s$ as the estimate for the former system and the simple arithmetic mean of the first-stage unit means as the estimate for the latter system.

For the system of sampling with probability proportional to size and a fixed number of second-stage units selected from each first-stage unit, we have seen that $\bar{z}_S$ provides an unbiased estimate of the population mean, and its mean square error is given by (14). This may be rewritten as

$$M.S.E.\ (\bar{z}_s) = \frac{1}{nN} \sum_{i=1}^{N} \frac{M_i}{\bar{M}} \left\{(\bar{y}_{i.} - \bar{y}_{..})^2 - \frac{S_i^2}{M_i}\right\} + \frac{1}{nmN} \sum_{i=1}^{N} \frac{M_i}{\bar{M}} S_i^2 \quad (73)$$

For convenience, we shall suppose that first-stage units of the same size are grouped together. Defining then ρ_i as the intra-class correlation within first-stage units of the same size, we may write

$$\begin{aligned}
\rho_i\sigma^2 &= E\{(y_{ij} - \bar{y}_{..})(y_{ik} - \bar{y}_{..}) \mid i\} \\
&= E\{(y_{ij} - \bar{y}_{i.} + \bar{y}_{i.} - \bar{y}_{..})(y_{ik} - \bar{y}_{i.} + \bar{y}_{i.} - \bar{y}_{..}) \mid i\} \\
&= E\{(y_{ij} - \bar{y}_{i.})(y_{ik} - \bar{y}_{i.}) \mid i\} + E\{(\bar{y}_{i.} - \bar{y}_{..})^2 \mid i\} \\
&= \frac{1}{M_i(M_i - 1)} \sum_{j \neq k = 1}^{M_i} (y_{ij} - \bar{y}_{i.})(y_{ik} - \bar{y}_{i.}) + (\bar{y}_{i.} - \bar{y}_{..})^2 \\
&= \frac{1}{M_i(M_i - 1)} \left[\left\{\sum_{j=1}^{M_i} (y_{ij} - \bar{y}_{i.})\right\}^2 - \sum_{j=1}^{M_i} (y_{ij} - \bar{y}_{i.})^2\right] \\
&\qquad + (\bar{y}_{i.} - \bar{y}_{..})^2 \\
&= -\frac{S_i^2}{M_i} + (\bar{y}_{i.} - \bar{y}_{..})^2 \qquad (74)
\end{aligned}$$

Substituting the result in (73), we obtain

$$M.S.E.\ (\bar{z}_s) = \frac{\sigma^2}{nN} \sum_{i=1}^{N} \frac{M_i}{\bar{M}} \rho_i + \frac{1}{nmN} \sum_{i=1}^{N} \frac{M_i}{\bar{M}} S_i^2 \qquad (75)$$

For a sub-sampling design involving the use of equal selection probabilities at both stages, the simple arithmetic mean estimate is known to be biased and the mean square error is derived directly from (84) and (85) of Chapter VII. Ignoring the finite multiplier at the first stage of sampling, this is given by

$$\begin{aligned} M.S.E.\,(\bar{y}_s) &= \frac{1}{nN}\sum_{i=1}^{N}\left\{(\bar{y}_{i.}-\bar{y}_{..})^2-\frac{\mathbf{S}_i^2}{M_i}\right\} \\ &\quad+\frac{1}{nmN}\sum_{i=1}^{N}\mathbf{S}_i^2+\left(1-\frac{1}{n}\right)(\bar{y}_{N.}-\bar{y}_{..})^2 \\ &= \frac{\sigma^2}{nN}\sum_{i=1}^{N}\rho_i+\frac{1}{nmN}\sum_{i=1}^{N}\mathbf{S}_i^2 \\ &\quad+\left(1-\frac{1}{n}\right)(\bar{y}_{N.}-\bar{y}_{..})^2 \qquad (76)\end{aligned}$$

The difference between the two mean square errors is, therefore, given by

$$\begin{aligned} M.S.E.\,(\bar{y}_s)-M.S.E.\,(\bar{z}_s) &= -\frac{\sigma^2}{nN}\sum_{i=1}^{N}\rho_i\left(\frac{M_i}{\bar{M}}-1\right) \\ &\quad-\frac{1}{nmN}\sum_{i=1}^{N}\mathbf{S}_i^2\left(\frac{M_i}{\bar{M}}-1\right) \\ &\quad+\left(1-\frac{1}{n}\right)(\bar{y}_{N.}-\bar{y}_{..})^2\end{aligned}$$

Putting $\mathbf{S}_i^2=\sigma^2(1-\delta_i)$ where δ_i has a meaning similar to ρ_i but is not necessarily equal to ρ_i, we may express the difference as

$$\begin{aligned} &= -\frac{\sigma^2}{nN}\sum_{i=1}^{N}\rho_i\left(\frac{M_i}{\bar{M}}-1\right) \\ &\quad+\frac{\sigma^2}{nmN}\sum_{i=1}^{N}\delta_i\left(\frac{M_i}{\bar{M}}-1\right) \\ &\quad+\left(1-\frac{1}{n}\right)(\bar{y}_{N.}-\bar{y}_{..})^2 \qquad (77)\end{aligned}$$

Now both ρ_i and δ_i will usually decrease as M_i increases, so that the covariance between ρ_i and $M_i/\bar{M}$ and between δ_i and $M_i/\bar{M}$ will be negative. The first term with its negative sign will, therefore, be positive but the second term will be negative but of a smaller order owing to the presence of the factor $1/m$, while the third term will always remain positive. We, therefore, conclude that $\bar{y}_S$ will ordinarily have a higher mean square error than $\bar{z}_S$, showing the superiority of $\bar{z}_S$ over the estimate $\bar{y}_S$.

8.9* Sub-Sampling without Replacement

In the sub-sampling procedure considered in the previous sections we have assumed that sub-samples from the same first-stage unit are drawn independently of each other. We shall now consider a procedure in which sub-sampling is carried out wholly without replacement, that is to say, that if any first-stage unit occurs γ times in the sample, a sub-sample of $m\gamma$ units will be drawn therefrom without replacement.

Following the previous notation, let

$$z_{ij} = \frac{M_i}{M_0 P_i} \; y_{ij}$$

and

$\bar{z}_{i,\,m\gamma_i}$ = the mean per second-stage unit of the sub-sample drawn from the i-th first-stage unit

Consider the estimate $\bar{z}_S'$, given by

$$\bar{z}_s' = \frac{1}{n} \sum_{i=1}^{N} \gamma_i \bar{z}_{i,\,m\gamma_i} \tag{78}$$

where γ_i is a random variable with possible values $0, 1, 2, \ldots, n$, such that $\sum\limits_{i=1}^{N} \gamma_i = n$, and the probability that γ_i is equal to r is given by the $(r + 1)$-th term of the binomial

$$\{P_i + (1 - P_i)\}^n$$

namely,

$$P\,\{\gamma_i = r\} = \binom{n}{r} P_i^r\,(1 - P_i)^{n-r}$$

It is easily shown that $\bar{z}_s'$ is an unbiased estimate of the population mean $\bar{y}_{..}$. For, we have

$$E\,(\bar{z}_s') = \frac{1}{n}\,E\left\{\sum_{i=1}^{N} \gamma_i \bar{z}_{i,\;m\gamma_i}\right\}$$

$$= \frac{1}{n}\,E\left\{\sum_{i=1}^{N} \gamma_i E\,(\bar{z}_{i,\,m\gamma_i} \mid i,\, m\gamma_i)\right\}$$

$$= \frac{1}{n}\,E\left\{\sum_{i=1}^{N} \gamma_i\, \bar{z}_{i.}\right\}$$

$$= \frac{1}{n} \sum_{i=1}^{N} E\,(\gamma_i)\, \bar{z}_{i.} \qquad (79)$$

We already know that for a binomial distribution

$$E\,(\gamma_i) = nP_i \qquad (80)$$

On substituting from (80) in (79), we get

$$E\,(\bar{z}_s') = \frac{1}{n} \sum_{i=1}^{N} nP_i \bar{z}_{i.}$$

$$= \sum_{i=1}^{N} P_i \bar{z}_{i.}$$

$$= \bar{z}_{..} \qquad (81)$$

$$= \sum_{i=1}^{N} \frac{M_i}{M_0}\; \bar{y}_{i.}$$

$$= \bar{y}_{..} \qquad (82)$$

To obtain the sampling variance of the estimate, we have

$$V(\bar{z}_s') = E\left\{\frac{1}{n}\sum_{i=1}^{N}\gamma_i\bar{z}_{i,\,m\gamma_i} - \bar{z}_{..}\right\}^2$$

$$= E\left\{\frac{1}{n}\sum_{i=1}^{N}\gamma_i\bar{z}_{i,\,m\gamma_i} - \frac{1}{n}\sum_{i=1}^{N}\gamma_i\bar{z}_{i.} + \frac{1}{n}\sum_{i=1}^{N}\gamma_i\bar{z}_{i.} - \bar{z}_{..}\right\}^2$$

$$= \frac{1}{n^2}\,E\left\{\sum_{i=1}^{N}\gamma_i\,(\bar{z}_{i,\,m\gamma_i} - \bar{z}_{i.} + \bar{z}_{i.} - \bar{z}_{..})\right\}^2$$

$$= \frac{1}{n^2}\,E\left\{\sum_{i=1}^{N}\gamma_i^2\,(\bar{z}_{i,\,m\gamma_i} - \bar{z}_{i.} + \bar{z}_{i.} - \bar{z}_{..})^2\right\}$$

$$+ \frac{1}{n^2}\,E\left\{\sum_{i\neq i'=1}^{N}\gamma_i\gamma_{i'}\,(\bar{z}_{i,\,m\gamma_i} - \bar{z}_{i.} + \bar{z}_{i.} - \bar{z}_{..})\right.$$

$$\left.\times\,(\bar{z}_{i',\,m\gamma_{i'}} - \bar{z}_{i'.} + \bar{z}_{i'.} - \bar{z}_{..})\right\} \qquad (83)$$

Taking the first term in (83), we write

$$E\left\{\sum_{i=1}^{N}\gamma_i^2\,(\bar{z}_{i,\,m\gamma_i} - \bar{z}_{i.} + \bar{z}_{i.} - \bar{z}_{..})^2\right\}$$

$$= E\left[\sum_{i=1}^{N}\gamma_i^2\,\{(\bar{z}_{i,\,m\gamma_i} - \bar{z}_{i.})^2 + (\bar{z}_{i.} - \bar{z}_{..})^2\right.$$

$$\left. + 2\,(\bar{z}_{i,\,m\gamma_i} - \bar{z}_{i.})\,(\bar{z}_{i.} - \bar{z}_{..})\}\right]$$

$$= E\left[\sum_{i=1}^{N}\gamma_i^2\left\{E\Big((\bar{z}_{i,\,m\gamma_i} - \bar{z}_{i.})^2 \mid i, \gamma_i\Big) + E\Big((\bar{z}_{i.} - \bar{z}_{..})^2 \mid i, \gamma_i\Big)\right.\right.$$

$$\left.\left. + 2E\Big((\bar{z}_{i,\,m\gamma_i} - \bar{z}_{i.})\,(\bar{z}_{i.} - \bar{z}_{..}) \mid i, \gamma_i\Big)\right\}\right]$$

$$= E\left\{\sum_{i=1}^{N}\gamma_i^2\left(\frac{1}{m\gamma_i} - \frac{1}{M_i}\right)\mathbf{S}_{iz}^2 + \sum_{i=1}^{N}\gamma_i^2\,(\bar{z}_{i.} - \bar{z}_{..})^2\right\} \qquad (84)$$

Likewise, the second term in (83) gives

$$E\left\{\sum_{i\neq i'=1}^{N}\gamma_i\gamma_{i'}(\bar{z}_{i,\,m\gamma_i}-\bar{z}_{i.}+\bar{z}_{i.}-\bar{z}_{..})(\bar{z}_{i',\,m\gamma_{i'}}-\bar{z}_{i'.}+\bar{z}_{i'.}-\bar{z}_{..})\right\}$$

$$=E\Bigg[\sum_{i\neq i'=1}^{N}\gamma_i\gamma_{i'}\{(\bar{z}_{i,\,m\gamma_i}-\bar{z}_{i.})\ (\bar{z}_{i',\,m\gamma_{i'}}-\bar{z}_{i'.})$$

$$+(\bar{z}_{i,m\gamma_i}-\bar{z}_{i.})(\bar{z}_{i'.}-\bar{z}_{..})+(\bar{z}_{i.}-\bar{z}_{..})(\bar{z}_{i',\,m\gamma_{i'}}-\bar{z}_{i'.})$$

$$+(\bar{z}_{i.}-\bar{z}_{..})\ (\bar{z}_{i'.}-\bar{z}_{..})\}\Bigg]$$

$$=E\left\{\sum_{i\neq i'=1}^{N}\gamma_i\gamma_{i'}\,(\bar{z}_{i.}-\bar{z}_{..})\ (\bar{z}_{i'.}-\bar{z}_{..})\right\} \tag{85}$$

since the expectation of the other three terms is clearly zero.

Substituting from (84) and (85) in (83), we may write

$$V(\bar{z}_{i}') = \frac{1}{n^2}\,E\Bigg[\sum_{i=1}^{N}\frac{\gamma_i\mathbf{S}_{iz}{}^2}{m}-\sum_{i=1}^{N}\gamma_i{}^2\,\frac{\mathbf{S}_{iz}{}^2}{M_i}$$

$$+\left\{\sum_{i=1}^{N}\gamma_i\ (\bar{z}_{i.}-\bar{z}_{..})\right\}^2\Bigg]$$

$$=\frac{1}{n^2}\Bigg[\left\{\sum_{i=1}^{N}\frac{E\,(\gamma_i)\mathbf{S}_{iz}{}^2}{m}\right\}-\left\{\sum_{i=1}^{N}\frac{E\ (\gamma_i{}^2)\,\mathbf{S}_{iz}{}^2}{M_i}\right\}$$

$$+\ E\left\{\sum_{i=1}^{N}\gamma_i\,(\bar{z}_{i.}-\bar{z}_{..})\right\}^2\Bigg] \tag{86}$$

From (80), we know that $E\,(\gamma_i)=nP_i$. To evaluate $E\,(\gamma_i{}^2)$, we write

$$E\,(\gamma_i{}^2)=\sum_{r=1}^{n}r^2\,P\,(\gamma_i=r)$$

$$=\sum_{r=1}^{n}r^2\binom{n}{r}\,P_i{}^r\,(1-P_i)^{n-r}$$

which is clearly the second moment of the binomial distribution, and is, therefore,

$$= nP_i(1 - P_i) + n^2P_i^2 \tag{87}$$

To evaluate the third term in (86), we require the values of $E(\gamma_i^2)$ and $E(\gamma_i\gamma_j)$. We write

$$\begin{aligned} E(\gamma_i\gamma_j) &= E\{E(\gamma_i\gamma_j \mid \gamma_i)\} \\ &= E\{\gamma_i E(\gamma_j \mid \gamma_i)\} \end{aligned} \tag{88}$$

where $E(\gamma_j \mid \gamma_i)$ denotes the expected value of γ_j, given γ_i. Now the probability of drawing γ_j, given γ_i, from a sample of $(n - \gamma_i)$ is $P_j/(1 - P_i)$. Hence, by analogy with (80),

$$E(\gamma_j \mid \gamma_i) = (n - \gamma_i) \cdot \frac{P_j}{1 - P_i} \tag{89}$$

Substituting from (89) in (88), we obtain

$$\begin{aligned} E(\gamma_i\gamma_j) &= E\left\{\gamma_i \cdot (n - \gamma_i) \cdot \frac{P_j}{1 - P_i}\right\} \\ &= \frac{nP_j}{1 - P_i} E(\gamma_i) - \frac{P_j}{1 - P_i} E(\gamma_i^2) \\ &= \frac{nP_j}{1 - P_i} \cdot nP_i - \frac{P_j}{1 - P_i}\{nP_i(1 - P_i) + n^2P_i^2\} \\ &= n(n - 1)P_iP_j \end{aligned} \tag{90}$$

Using (87) and (90), the third term in (86) may now be written as

$$\begin{aligned} E\left\{\sum_{i=1}^{N} \gamma_i(\bar{z}_{i.} - \bar{z}_{..})\right\}^2 &= \sum_{i=1}^{N} E(\gamma_i^2)(\bar{z}_{i.} - \bar{z}_{..})^2 \\ &\quad + \sum_{i \neq j=1}^{N} E(\gamma_i\gamma_j)(\bar{z}_{i.} - \bar{z}_{..})(\bar{z}_{j.} - \bar{z}_{..}) \\ &= \sum_{i=1}^{N} \{nP_i(1 - P_i) + n^2P_i^2\}(\bar{z}_{i.} - \bar{z}_{..})^2 \\ &\quad + \sum_{i \neq j=1}^{N} \{n(n - 1)P_iP_j\} \\ &\quad \times (\bar{z}_{i.} - \bar{z}_{..})(\bar{z}_{j.} - \bar{z}_{..}) \end{aligned}$$

$$
\begin{aligned}
&= n \sum_{i=1}^{N} P_i (\bar{z}_{i.} - \bar{z}_{..})^2 \\
&\qquad + n(n-1) \left\{ \sum_{i=1}^{N} P_i (\bar{z}_{i.} - \bar{z}_{..}) \right\}^2 \\
&= n\sigma_{bz}^2 \qquad (91)
\end{aligned}
$$

since $\sum\limits_{i=1}^{N} P_i (\bar{z}_{i.} - \bar{z}_{..})$ is clearly zero.

Hence, substituting from (80), (87) and (91) in (86), we obtain on rearranging terms,

$$
\begin{aligned}
V(\bar{z}_s') = \frac{\sigma_{bz}^2}{n} + \frac{1}{n} \sum_{i=1}^{N} P_i \left(\frac{1}{m} - \frac{1}{M_i} \right) S_{iz}^2 \\
- \frac{n-1}{n} \sum_{i=1}^{N} P_i^2 \frac{S_{iz}^2}{M_i} \qquad (92)
\end{aligned}
$$

which can also be written as

$$
\begin{aligned}
V(\bar{z}_s') = \frac{1}{n} \left\{ \sum_{i=1}^{N} \frac{M_i^2 \bar{y}_{i.}^2}{M_0^2 P_i} - \bar{y}_{..}^2 \right\} \\
+ \frac{1}{n} \sum_{i=1}^{N} \frac{M_i^2}{M_0^2 P_i} \left(\frac{1}{m} - \frac{1}{M_i} \right) S_i^2 \\
- \frac{n-1}{n} \sum_{i=1}^{N} \frac{M_i}{M_0^2} S_i^2 \qquad (93)
\end{aligned}
$$

When $P_i = M_i / M_0$, (93) reduces to

$$
\begin{aligned}
V(\bar{z}_s') = \frac{1}{n} \sum_{i=1}^{N} \frac{M_i}{M_0} (\bar{y}_{i.} - \bar{y}_{..})^2 \\
+ \frac{1}{n} \sum_{i=1}^{N} \frac{M_i}{M_0} \left(\frac{1}{m} - \frac{1}{M_i} \right) S_i^2 \\
- \frac{n-1}{nM_0} \sum_{i=1}^{N} \frac{M_i}{M_0} S_i^2 \qquad (94)
\end{aligned}
$$

If further $\mathbf{S}_i^2 = \mathbf{S}_w^2$, we have

$$V(\bar{z}_s') = \frac{\sigma_b^2}{n} + \frac{\mathbf{S}_w^2}{n}\left(\frac{1}{m} - \frac{1}{\bar{M}}\right) - \frac{n-1}{n} \cdot \frac{\mathbf{S}_w^2}{N\bar{M}} \tag{95}$$

It is interesting to compare (93) with (12) after putting $m_i = m$ in the latter. The comparison shows that the variance is reduced by

$$\frac{n-1}{n} \sum_{i=1}^{N} \frac{M_i}{M_0^2} \mathbf{S}_i^2$$

showing that the procedure of sub-sampling considered in this section is more efficient than the previous procedure. This is in accordance with expectation. The gain is however small, since the contribution to the variance from any modification at the sub-sampling stage must necessarily be of a second order (Sukhatme and Narain, 1952).

8.10* Estimation of the Variance from the Sample when Sub-Sampling is carried out without Replacement

Consider the mean square between the first-stage unit means in the sample defined by

$$\begin{aligned}(n-1)\, s_{bz}'^2 &= \sum_{i=1}^{N} \gamma_i \left(\bar{z}_{i,\, m\gamma_i} - \bar{z}_s'\right)^2 \\ &= \sum_{i=1}^{N} \gamma_i \bar{z}^2{}_{i,\, m\gamma_i} - n\bar{z}_s'^2 \end{aligned} \tag{96}$$

Taking expectations, we obtain

$$\begin{aligned}(n-1)\, E\left(s_{bz}'^2\right) &= E\left\{\sum_{i=1}^{N} \gamma_i E\left(\bar{z}^2{}_{i,\, m\gamma_i} \mid i, \gamma_i\right)\right\} - nE\left(\bar{z}_s'^2\right) \\ &= E\left[\sum_{i=1}^{N} \gamma_i \left\{\bar{z}_{i.}^2 + \left(\frac{1}{m\gamma_i} - \frac{1}{M_i}\right) \mathbf{S}_{iz}^2\right\}\right] \\ &\qquad - nE\left(\bar{z}_s'^2\right) \\ &= E\left[\sum_{i=1}^{N} \gamma_i \left\{\bar{z}_{i.}^2 - \frac{\mathbf{S}_{iz}^2}{M_i}\right\}\right] + E\left\{\sum^{n'} \frac{\mathbf{S}_{iz}^2}{m}\right\} \\ &\qquad - nE\left(\bar{z}_s'^2\right)\end{aligned}$$

where n' denotes the number of different first-stage units included in the sample. Substituting from (92) for $E\,(\bar{z}_S'^2)$, we may write

$$
\begin{aligned}
(n-1)\,E\,(s_{bz}'^2) = n \sum_{i=1}^{N} P_i \left\{\bar{z}_{i.}^2 - \frac{\mathrm{S}_{iz}^2}{M_i}\right\} + E\left\{\sum^{n} \frac{\mathrm{S}_{iz}^2}{m\gamma_i}\right\} \\
- n\left[\bar{z}_{..}^2 + \frac{\sigma_{bz}^2}{n} + \frac{1}{n}\sum_{i=1}^{N} P_i \left(\frac{1}{m} - \frac{1}{M_i}\right) \mathrm{S}_{iz}^2\right. \\
\left. - \frac{n-1}{n}\sum_{i=1}^{N} P_i^2 \frac{\mathrm{S}_{iz}^2}{M_i}\right]
\end{aligned}
$$

or, on dividing by $(n-1)$, we have

$$
\begin{aligned}
E\,(s_{bz}'^2) = \sigma_{bz}^2 - \frac{n}{n-1}\sum_{i=1}^{N} P_i \frac{\mathrm{S}_{iz}^2}{M_i} + \frac{1}{n-1}\,E\left\{\sum^{n} \frac{\mathrm{S}_{iz}^2}{m\gamma_i}\right\} \\
- \frac{1}{n-1}\sum_{i=1}^{N} P_i \left(\frac{1}{m} - \frac{1}{M_i}\right)\mathrm{S}_{iz}^2 \\
+ \sum_{i=1}^{N} P_i^2 \frac{\mathrm{S}_{iz}^2}{M_i} \qquad (97)
\end{aligned}
$$

Hence

$$
\begin{aligned}
s_{bz}'^2 = \hat{\sigma}_{bz}^2 - \frac{1}{n-1}\sum^{n} \frac{\hat{\mathrm{S}}_{iz}^2}{M_i} + \frac{1}{n-1}\sum^{n} \frac{\hat{\mathrm{S}}_{iz}^2}{m\gamma_i} \\
- \frac{1}{n\,(n-1)}\sum^{n} \left(\frac{1}{m} - \frac{1}{M_i}\right)\hat{\mathrm{S}}_{iz}^2 \\
+ \frac{1}{n}\sum^{n} P_i \frac{\hat{\mathrm{S}}_{iz}^2}{M_i} \qquad (98)
\end{aligned}
$$

or

$$\text{Est. } \sigma_{bz}^2 = s_{bz}'^2 - \frac{1}{n-1} \sum^{n} \left(\frac{1}{m\gamma_i} - \frac{1}{M_i} \right) s_{iz}'^2$$

$$+ \frac{1}{n(n-1)} \sum^{n} \left(\frac{1}{m} - \frac{1}{M_i} \right) s_{iz}'^2$$

$$- \frac{1}{n} \sum^{n} P_i \frac{s_{iz}'^2}{M_i} \qquad (99)$$

where

$$s_{iz}'^2 = \frac{\sum^{m\gamma_i} (z_{ij} - \bar{z}_{i,\, m\gamma_i})^2}{m\gamma_i - 1}$$

If $\mathbf{S}_{iz}^2 = \mathbf{S}_{wz}^2$, (97) is simplified, being given by

$$E(s_{bz}'^2) = \sigma_{bz}^2 - \mathbf{S}_{wz}^2 \sum_{i=1}^{N} \frac{P_i}{M_i} + \frac{\mathbf{S}_{wz}^2}{m} \cdot \frac{E(n') - 1}{n - 1}$$

$$+ \mathbf{S}_{wz}^2 \sum_{i=1}^{N} \frac{P_i^2}{M_i}$$

whence, we get

$$\text{Est. } \sigma_{bz}^2 = s_{bz}'^2 - \frac{E(n') - 1}{n - 1} \cdot \frac{s_{wz}'^2}{m} + \frac{s_{wz}'^2}{\bar{M}_h} - \frac{s_{wz}'^2}{n} \sum^{n} \frac{P_i}{M_i} \qquad (100)$$

where

$$s_{wz}'^2 = \frac{\sum^{n'} \sum^{m\gamma_i} (z_{ij} - \bar{z}_{i,\, m\gamma_i})^2}{nm - n'}$$

and $\bar{M}_h$ denotes the harmonic mean of M_i's in the sample. Further, when $P_i = M_i / N\bar{M}$, we have

$$\text{Est. } \sigma_b^2 = s_b'^2 - s_w'^2 \left\{ \frac{E(n') - 1}{m(n-1)} - \frac{1}{\bar{M}} + \frac{1}{N\bar{M}} \right\} \qquad (101)$$

Substituting from (99) in the expression (92) for the variance of $\bar{z}_s'$, we get

$$\text{Est. } V(\bar{z}_s') = \frac{s_{bz}'^2}{n} + \frac{1}{mn(n-1)} \sum^{n} \left(1 - \frac{1}{\gamma_i}\right) s_{iz}'^2 - \frac{1}{n} \sum^{n} P_i \frac{s_{iz}'^2}{M_i} \tag{102}$$

When $P_i = M_i/M_0$, and $s_{iz}'^2$ is identical with $s_i'^2$ and can be replaced by a pooled estimate $s_w'^2$, we get

$$\text{Est. } V(\bar{z}_s') = \frac{s_b'^2}{n} - \frac{s_w'^2}{nm} \left\{ \frac{E(n') - 1}{n - 1} - 1 + \frac{nm}{N\bar{M}} \right\} \tag{103}$$

The sampling procedure described in this section and the preceding one is widely used in India for estimating the acreage under crops. The design is particularly suitable for the introduction of the improved methods of estimating crop acreages in tracts which are cadastrally surveyed. Thus, in Orissa, in India, where this design was first used, a village is used as the first-stage unit of sampling and selected with probability proportional to the number of survey numbers (fields) in the village. Each selected village is further divided into sub-units of 8 consecutive survey numbers, 1–8, 9–16, 17–24, etc., the last sub-unit consisting of the remainder. From the sub-units so formed, 4 sub-units are selected, giving an equal chance to all sub-units in the village. If a village occurs more than once in the sample, say γ times, a sub-sample of 4γ sub-units is selected from it.

The design derives its efficiency from two factors:

(i) the high correlation between the number of survey numbers in a village and the crop acreage, and

(ii) the convenience and economy in field work arising from the choice of natural units as the sampling units at each stage.

Uncultivated land, such as that occupied by dwellings, lying barren, or used as grassland, is usually given a single survey number in India, while the cultivated land, which is divided into a large number of fields, is given at least as many survey numbers as the number of holders in the village. Compared to the design which makes use of artificial units like square grids marked with the help of latitude and longitude on the map, such as for example the one used in the Bihar and the Bengal surveys (Mahalanobis, 1945 and 1948), and in which apart from administrative inconvenience in locating the unit on the ground, a large proportion of units falls into uncultivated tracts, this design is found to be not only convenient for field work, but also statistically efficient. For further reference the reader is referred to the report on the estimation of acreage under crops in Orissa (I.C.A.R., 1950).

8.11* Stratification and the Gain Due to it

In this section we shall give the formulæ for the estimate of the population mean in stratified sampling and its variance, and then proceed to estimate from a stratified sample the change in variance due to stratification.

Let P_{ti} denote the selection probability assigned to the i-th first-stage unit within the t-th stratum, so that

$$\sum_{i=1}^{N_t} P_{ti} = 1 \qquad (t = 1, 2, \ldots, k)$$

Let n_t denote the number of first-stage units to be included in the sample from the t-th stratum, so that

$$\sum_{t=1}^{k} n_t = n$$

and m_{ti} the number of second-stage units to be selected from the i-th first-stage unit of the t-th stratum. Defining then

$$z_{tij} = \frac{M_{ti}}{M_{t0}} \cdot \frac{1}{P_{ti}} y_{tij} \qquad \begin{matrix} (i = 1, 2, \ldots, n_t) \\ (j = 1, 2, \ldots, m_{ti}) \end{matrix} \tag{104}$$

it is easy to see that

$$\bar{z}_{ts} = \frac{1}{n_t} \sum^{n_t} \bar{z}_{ti(m_{ti})} \tag{105}$$

provides an unbiased estimate of the population mean of the t-th stratum, and

$$\bar{z}_w = \sum_{t=1}^{k} \lambda_t \, \bar{z}_{ts} \tag{106}$$

that of the mean for the whole population, where

$$\lambda_t = \frac{M_{t0}}{M_0} \tag{107}$$

The variance of $\bar{z}_w$ is given by

$$V(\bar{z}_w) = \sum_{t=1}^{k} \lambda_t^2 \left\{ \frac{\sigma^2_{tb(zt)}}{n_t} + \frac{1}{n_t} \sum_{i=1}^{N_t} P_{ti} \left(\frac{1}{m_{ti}} - \frac{1}{M_{ti}} \right) S^2_{ti(zt)} \right\} \tag{108}$$

where

$$\sigma^2_{tb(zt)} = \sum_{i=1}^{N_t}{}' P_{ti} \, (\bar{z}_{ti.} - \bar{z}_{t..})^2 \tag{109}$$

and

$$S^2_{ti(zt)} = \frac{1}{M_{ti} - 1} \sum_{j=1}^{M_{ti}} (z_{tij} - \bar{z}_{ti.})^2 \tag{110}$$

If the sample were chosen as an unstratified sample, then the population mean would be estimated by

$$\bar{z}_s = \frac{1}{n} \sum^{n} \bar{z}_{i(m_i)} \tag{111}$$

where

$$z_{ij} = \frac{M_i}{M_0} \cdot \frac{1}{P_i} \, y_{ij} \tag{112}$$

with its variance given by

$$V(\bar{z}_s) = \frac{\sigma^2_{b(z)}}{n} + \frac{1}{n} \sum_{i=1}^{N} P_i \left(\frac{1}{m_i} - \frac{1}{M_i} \right) S^2_{i(z)} \tag{113}$$

where

$$\sigma^2_{b(z)} = \sum_{i=1}^{N} P_i \, (\bar{z}_{i.} - \bar{z}_{..})^2 \tag{114}$$

and

$$\mathbf{S}^2{}_{l(z)} = \frac{1}{M_l - 1} \sum_{j=1}^{M_l} (z_{lj} - \bar{z}_{l.})^2 \tag{115}$$

The difference between (113) and (108) gives the change in variance due to stratification. To estimate it from a stratified sample we shall suppose that the l-th unit in the population corresponds to the i-th unit in the t-th stratum, so that

$$P_{ti} = \frac{P_l}{P_{t.}} \tag{116}$$

where

$$P_{t.} = \overset{N_t}{\Sigma} P_i$$

and rewrite equation (113) in a form more suitable for estimating the gain due to stratification. Substituting for P_l from (116) in (114), and noting that $z_{lj} = z_{tij}\,(\lambda_t/P_{t.})$, we write

$$\sigma^2{}_{b(z)} = \sum_{t=1}^{k} P_{t.} \sum_{i=1}^{N_t} P_{ti} \left(\bar{z}_{ti.}\,\frac{\lambda_t}{P_{t.}} - \bar{z}_{..}\right)^2$$

$$= \sum_{t=1}^{k} P_{t.} \sum_{i=1}^{N_t} P_{ti} \left(\bar{z}_{ti.}\,\frac{\lambda_t}{P_{t.}} - \bar{z}_{t..}\,\frac{\lambda_t}{P_{t.}} + \bar{z}_{t..}\,\frac{\lambda_t}{P_{t.}} - \bar{z}_{..}\right)^2$$

$$= \sum_{t=1}^{k} P_{t.} \sum_{i=1}^{N_t} P_{ti} \left\{\frac{\lambda_t^2}{P_{t.}^2}\,(\bar{z}_{ti.} - \bar{z}_{t..})^2 + \left(\frac{\lambda_t}{P_{t.}}\,\bar{z}_{t..} - \bar{z}_{..}\right)^2\right\}$$

$$= \sum_{t=1}^{k} \frac{\lambda_t^2}{P_{t.}}\,\sigma^2{}_{tb(zt)} + \sum_{t=1}^{k} \frac{\lambda_t^2}{P_{t.}}\,\bar{z}_{t..}{}^2 - \bar{z}_{..}{}^2 \tag{117}$$

Also

$$P_l\mathbf{S}^2{}_{l(z)} = \frac{\lambda_t^2}{P_{t.}}\,P_{ti}\mathbf{S}^2{}_{ti(zt)} \tag{118}$$

Hence substituting for $\sigma^2{}_{b(z)}$ from (117) and for $P_l\mathbf{S}^2{}_{l(z)}$ from (118) in (113), we get

$$V(\bar{z}_s) = \frac{1}{n}\sum_{t=1}^{k}\frac{\lambda_t{}^2}{P_{t.}}\sigma^2{}_{tb(zt)} + \frac{1}{n}\left\{\sum_{t=1}^{k}\frac{\lambda_t{}^2}{P_{t.}}\ \bar{z}_{t..}{}^2 - \bar{z}_{..}{}^2\right\}$$

$$+\frac{1}{n}\sum_{t=1}^{k}\frac{\lambda_t{}^2}{P_{t.}}\sum_{i=1}^{N_t}P_{ti}\left(\frac{1}{m_{ti}}-\frac{1}{M_{ti}}\right)\mathbf{S}^2{}_{ti(zt)} \tag{119}$$

The change in variance due to stratification is thus given by the difference of (119) and (108), namely,

$$\left\{V_{US}-V_S\right\} = \sum_{t=1}^{k}\lambda_t{}^2\left(\frac{1}{nP_{t.}}-\frac{1}{n_t}\right)\sigma^2{}_{tb(zt)}$$

$$+\frac{1}{n}\left\{\sum_{t=1}^{k}\frac{\lambda_t{}^2}{P_{t.}}\ \bar{z}_{t..}{}^2-\bar{z}_{..}{}^2\right\}$$

$$+\sum_{t=1}^{k}\lambda_t{}^2\left(\frac{1}{nP_{t.}}-\frac{1}{n_t}\right)$$

$$\times\sum_{i=1}^{N_t}P_{ti}\left(\frac{1}{m_{ti}}-\frac{1}{M_{ti}}\right)\mathbf{S}^2{}_{ti(zt)} \tag{120}$$

and it is this which we are required to estimate from a given stratified sample. We write

$$\text{Est. }\{V_{US}-V_S\} = \sum_{t=1}^{k}\lambda_t{}^2\left(\frac{1}{nP_{t.}}-\frac{1}{n_t}\right)\hat{\sigma}^2{}_{tb(zt)}$$

$$+\frac{1}{n}\left\{\sum_{t=1}^{k}\frac{\lambda_t{}^2}{P_{t.}}\text{ Est. }(\bar{z}_{t..}{}^2)-\text{Est. }(\bar{z}_{..}{}^2)\right\}$$

$$+\sum_{t=1}^{k}\frac{\lambda_t{}^2}{n_t}\left(\frac{1}{nP_{t.}}-\frac{1}{n_t}\right)\sum^{n_t}\left(\frac{1}{m_{ti}}-\frac{1}{M_{ti}}\right)$$

$$\times\hat{\mathbf{S}}^2{}_{ti(zt)} \tag{121}$$

Now, from (18), we have

$$\hat{\sigma}^2_{tb(zt)} = s^2_{tb(zt)} - \frac{1}{n_t} \sum^{n_t} \left(\frac{1}{m_{ti}} - \frac{1}{M_{ti}}\right) s^2_{ti(zt)} \tag{122}$$

$$= \frac{1}{n_t - 1} \sum^{n_t} \left\{\bar{z}_{ti(mti)} - \bar{z}_{ts}\right\}^2$$

$$- \frac{1}{n_t} \sum^{n_t} \left(\frac{1}{m_{ti}} - \frac{1}{M_{ti}}\right) \frac{1}{m_{ti} - 1} \sum^{m_{ti}} \left\{ z_{tij} - \bar{z}_{ti(mti)} \right\}^2$$

Also

$$\hat{\mathrm{S}}^2_{ti(zt)} = s^2_{ti(zt)} \tag{123}$$

It only remains to evaluate the middle terms in (121). We have

$$V(\bar{z}_{ts}) = E(\bar{z}_{ts}{}^2) - \bar{z}_{t..}{}^2$$

$$= \frac{\sigma^2_{tb(zt)}}{n_t} + \frac{1}{n_t} \sum_{i=1}^{Nt} P_{ti} \left(\frac{1}{m_{ti}} - \frac{1}{M_{ti}}\right) \mathrm{S}^2_{ti(zt)}$$

so that

$$\text{Est. } (\bar{z}_{t..}{}^2) = \bar{z}_{ts}{}^2 - \frac{\hat{\sigma}^2_{tb(zt)}}{n_t} - \frac{1}{n_t^2} \sum^{n_t} \left(\frac{1}{m_{ti}} - \frac{1}{M_{ti}}\right) \hat{\mathrm{S}}^2_{ti(zt)} \tag{124}$$

Similarly

$$\bar{z}_{..}{}^2 = E(\bar{z}_{w}{}^2) - V(\bar{z}_{w})$$

whence

$$\text{Est. } (\bar{z}_{..}{}^2) = \bar{z}_{w}{}^2 - \sum_{t=1}^{k} \lambda_t{}^2 \left\{\frac{\hat{\sigma}^2_{tb(zt)}}{n_t}\right.$$

$$\left. + \frac{1}{n_t^2} \sum^{n_t} \left(\frac{1}{m_{ti}} - \frac{1}{M_{ti}}\right) \hat{\mathrm{S}}^2_{ti(zt)} \right\} \tag{125}$$

Hence, from (124) and (125), we obtain

$$\text{Est.}\left\{\sum_{t=1}^{k}\frac{\lambda_t^2}{P_{t.}}\bar{z}_{t..}^2-\bar{z}_{..}^2\right\}=\sum_{t=1}^{k}\frac{\lambda_t^2}{P_{t.}}\bar{z}_{ts}^2-\bar{z}_{w}^2$$

$$-\sum_{t=1}^{k}\frac{\lambda_t^2}{n_t}\left(\frac{1}{P_{t.}}-1\right)\hat{\sigma}^2_{tb(zt)}$$

$$-\sum_{t=1}^{k}\frac{\lambda_t^2}{n_t^2}\left(\frac{1}{P_{t.}}-1\right)$$

$$\times\sum^{n_t}\left(\frac{1}{m_{ti}}-\frac{1}{M_{ti}}\right)\hat{\text{S}}^2_{ti(zt)} \tag{126}$$

Substituting from (122), (123) and (126) in (121), and collecting terms, we get

$$\text{Est.}\{V_{US}-V_S\}=\frac{1}{n}\left\{\sum_{t=1}^{k}\frac{\lambda_t^2}{P_{t.}}\bar{z}_{ts}^2-\bar{z}_{w}^2\right\}$$

$$+\sum_{t=1}^{k}\frac{\lambda_t^2}{n_t}\left(\frac{n_t}{n\,P_{t.}}-1-\frac{1}{nP_{t.}}+\frac{1}{n}\right)\hat{\sigma}^2_{tb(zt)}$$

$$+\sum_{t=1}^{k}\frac{\lambda_t^2}{n_t^2}\left(\frac{n_t}{nP_{t.}}-1-\frac{1}{nP_{t.}}+\frac{1}{n}\right)$$

$$\times\sum^{n_t}\left(\frac{1}{m_{ti}}-\frac{1}{M_{ti}}\right)\hat{\text{S}}^2_{ti(zt)} \tag{127}$$

On using (122), equation (127) is further simplified, being given by

$$\text{Est.}\{V_{US}-V_S\}=\frac{1}{n}\left\{\sum_{t=1}^{k}\frac{\lambda_t^2}{P_{t.}}\bar{z}_{ts}^2-\bar{z}_{w}^2\right\}$$

$$+\sum_{t=1}^{k}\frac{\lambda_t^2}{n_t}\left(\frac{n_t}{nP_{t.}}-1-\frac{1}{nP_{t.}}+\frac{1}{n}\right)s^2_{tb(zt)} \tag{128}$$

If $P_l = M_l/M_0$, then

$$P_{ti} = \frac{P_l}{\sum\limits^{N_t} P_l} = \frac{M_l}{\sum\limits^{N_t} M_l} = \frac{M_{ti}}{M_{t0}}$$

It follows immediately that

$$z_{lj} = y_{lj} = y_{tij} = z_{tij}$$

Also

$$P_{t.} = \sum^{N_t} P_l = \sum^{N_t} \frac{M_l}{M_0} = \frac{M_{t0}}{M_0} = \lambda_t$$

Then (128) reduces to

$$\text{Est.}\,\{V_{US} - V_S\}_{\,(P_{t\cdot} = \lambda_t)} = \frac{1}{n}\left\{\sum_{t=1}^{k} \lambda_t\, \bar{\bar{z}}_{t*}{}^2 - \bar{\bar{z}}_{..}{}^2\right\}$$
$$+ \sum_{t=1}^{k} \frac{\lambda_t}{nn_t}\left\{(n_t - 1) - \lambda_t\,(n-1)\right\} s^2_{tb(st)} \tag{129}$$

Example 8.1

A sample survey for estimating the acreage under paddy was carried out in Orissa State during 1950–51. Each district of the State was divided into a suitable number of strata by grouping together adjoining administrative divisions in the district. From each stratum a sample of villages was selected with probability proportional to the number of survey numbers in the village. Each selected village was divided into clusters of 8 consecutive survey numbers, 1–8, 9–16, 17–24, etc., the last cluster consisting of the remainder. From the clusters so formed a simple random sample of 4 clusters was drawn (without replacement). If, however, a village occurred more than once in the sample, say γ times, a sample of 4γ clusters was selected from it.

Table 8.1 shows the number of villages and the number of clusters in the population, the number of villages in the sample and the number n_t' of distinct villages in the sample, the estimated area under paddy per cluster and the values for $s_{tb}{}^2$, $s_{tw}{}^2$, $s_{tb}'^2$ and $s_{tw}'^2$ for each stratum. Calculate the sampling

TABLE 8.1

Area Survey on Paddy, Orissa, 1950–51

Values of Means, Mean Squares between Village Means (s_{tb}^2, $s_{tb}'^2$) and Mean Squares Within Villages (s_{tw}^2, $s_{tw}'^2$) in Acres per Cluster Basis

Stratum	N_t	$N_t\bar{M}_t$	n_t	n_t'	$\bar{y}_{tn_tm}$	s_{tb}^2	s_{tw}^2	$s_{tb}'^2$	$s_{tw}'^2$
1	434	71670	19	19	1·291	0·5058	1·0726	0·5058	1·0726
2	405	44114	13	12	2·597	3·7764	8·1920	3·7724	7·9921
3	565	33107	23	23	2·078	4·1255	9·0720	4·1255	9·0720
4	851	93734	34	32	2·098	1·8717	4·4411	1·8105	4·4334
5	271	24631	14	12	2·552	4·9885	12·3426	4·7173	12·1021
6	471	51776	18	18	1·675	0·6760	3·6231	0·6760	3·6231
7	347	44028	15	14	2·100	1·6487	3·3946	1·6482	3·3214
Sums	3344	363060	136	130					

variance of the estimate of the area under paddy per cluster in the district, assuming

(*a*) Sub-samples of 4 within each selected village were selected independently;

(*b*) A single sample of size 4γ was selected from each selected village (the method actually adopted).

Assume M_i/M_0P_i to be 1.

If the sampling variance of the estimated acreage can be supposed to be calculated under assumption (*a*), estimate the loss of efficiency which would have resulted from an unstratified sample.

The variance of the estimate of the area under paddy per cluster in the *t*-th stratum appropriate for assumption (*a*) is given by equation (19) as

$$\text{Est. } V(\bar{y}_{t\cdot}) = \frac{s_{tb}^2}{n_t} .$$

The variance of the district estimate is, therefore, given by

$$\text{Est. } V(\bar{y}_w) = \sum_{t=1}^{7} \lambda_t^2 \frac{s_{tb}^2}{n_t}$$

where

$$\lambda_t = \frac{N_t \bar{M}_t}{N\bar{M}} = \frac{M_{t0}}{M_0}$$

The weights λ_t are computed in the first column of Table 8.2. Substituting from the table, we have

$$\begin{aligned}\text{Est. } V(\bar{y}_w) &= (0{\cdot}19741)^2\ \frac{0{\cdot}5058}{19} + (0{\cdot}12151)^2\ \frac{3{\cdot}7764}{13} \\ &\quad + (0{\cdot}091189)^2\ \frac{4{\cdot}1255}{23} + (0{\cdot}25818)^2\ \frac{1{\cdot}8717}{34} \\ &\quad + (0{\cdot}067843)^2\ \frac{4{\cdot}9885}{14} + (0{\cdot}14261)^2\ \frac{0{\cdot}6760}{18} \\ &\quad + (0{\cdot}12127)^2\ \frac{1{\cdot}6487}{15} \\ &= 0{\cdot}001037 + 0{\cdot}004289 + 0{\cdot}001492 + 0{\cdot}003669 \\ &\quad + 0{\cdot}001640 + 0{\cdot}000764 + 0{\cdot}001616 \\ &= 0{\cdot}01451\end{aligned}$$

TABLE 8.2

Computations for Estimating Gain in Precision due to Stratification

Stratum t	$\lambda_t = \frac{N_t\bar{M}_t}{N\bar{M}}$	$\lambda_t \bar{y}_{tn_t m}$	$\lambda_t \bar{y}^2_{tn_t m}$	$\frac{\lambda_t}{nn_t}$	$n_t - 1 - (n-1)\lambda_t$	$(4) \times (5)\, s_{tb}^2$
	(1)	(2)	(3)	(4)	(5)	(6)
1	0·19741	0·2549	0·3290	·000076397	−8·6504	−·0003343
2	0·12151	0·3156	0·8195	·000068727	−4·4038	−·0011430
3	0·091189	0·1895	0·3938	·000029152	+9·6895	+·0011653
4	0·25818	0·5417	1·1364	·000055835	−1·8543	−·0001938
5	0·067843	0·1731	0·4418	·000035632	+3·8412	+·0006828
6	0·14261	0·2389	0·4001	·000058256	−2·2524	−·0000887
7	0·12127	0·2547	0·5348	·000059446	−2·3714	−·0002324
Sums		1·9684	4·0554			−·0001441

Under assumption (*b*), an estimate of the variance in a single stratum is given by equation (103). We may write

$$\text{Est. } V'(\bar{y}_{ts}) = \frac{s_{tb}'^2}{n_t} + \left(\frac{1}{n_t m} \cdot \frac{n_t - n_t'}{n_t - 1} - \frac{1}{N_t \bar{M}_t}\right) s_{tw}'^2$$

For example,

$$\text{Est. } V'(\bar{y}_{4s}) = \frac{1\cdot 8105}{34} + \left(\frac{1}{(34)\,(4)} \cdot \frac{2}{33} - \frac{1}{93734}\right)(4\cdot 4334)$$

$$= 0\cdot 053250 + (\cdot 00044563 - \cdot 00001067)\,(4\cdot 4334)$$

$$= 0\cdot 055178$$

The variance for the district estimate is then estimated by

$$\text{Est. } V'(\bar{y}_w) = \sum_{t=1}^{7} \lambda_t^2\, V'(\bar{y}_{ts})$$

$$= (0\cdot 19741)^2 \;\; (0\cdot 026606) \;\; + (0\cdot 12151)^2\,(0\cdot 30281)$$

$$+ (0\cdot 091189)^2\,(0\cdot 17910) + (0\cdot 25818)^2\,(0\cdot 055178)$$

$$+ (0\cdot 067843)^2\,(0\cdot 36971) + (0\cdot 14261)^2\,(0\cdot 037486)$$

$$+ (0\cdot 12127)^2 \;\; (0\cdot 11376)$$

$$= 0\cdot 01481$$

We note that contrary to expectation this value is larger than the first estimate, a fact which may be attributed to sampling errors in the estimates of σ_{tb}^2, $\sigma_{tb}'^2$, S_{tw}^2 and $S_{tw}'^2$; but the difference is negligible.

The difference between the variance of the mean of a stratified sample on assumption (*a*), and that of the mean of an unstratified sample is estimated by equation (129). The necessary computations are made in Table 8.2. We have

$$\text{Est. } \{V_{US} - V_S\} = \frac{1}{n}\left\{\sum_{t=1}^{k} \lambda_t\, \bar{z}_{ts}^2 - \bar{z}_w^2\right\}$$

$$+ \sum_{t=1}^{k} \frac{\lambda_t}{n n_t}\, \{(n_t - 1) - \lambda_t\,(n - 1)\}\, s^2{}_{tb(zt)}$$

$$= \frac{1}{136}\{4\cdot0554 - (1\cdot968)^2\} + (-0\cdot0001441)$$

$$= 0\cdot001341 - 0\cdot000144$$

$$= 0\cdot001197$$

Thus the relative increase in variance if the sample were not stratified would be

$$= \frac{0\cdot001197}{0\cdot01451}$$

$$= 0\cdot082$$

or

$$8\cdot2\%$$

8.12* Collapsed Strata

Hansen and Hurwitz (1943) advocate stratification to a degree where only one first-stage unit is selected from each stratum. Stratification to this degree may secure an optimum distribution of the sample when the strata are about equal, but offers one disadvantage in that an unbiased estimate of the error variance cannot be made. To overcome this difficulty Hansen and Hurwitz pool the strata in pairs which resemble each other as closely as possible and calculate an upper bound to the error variance of the estimate. One such procedure of calculating an upper bound will be described in this section.

We shall suppose that the population consists of only two strata and that from each stratum one first-stage unit is selected with probability proportional to the number of second-stage units in the first-stage unit. Let

N_1 and N_2 denote the total numbers of first-stage units in the first and second strata, respectively;

M_{1i} and M_{2j} the numbers of second-stage units in the i-th and the j-th first-stage units of the two strata, respectively ($i = 1, 2, \ldots, N_1$), ($j = 1, 2, \ldots, N_2$);

M_{10} and M_{20} — the total numbers of second-stage units in the first and the second strata, respectively, so that

$$M_0 = M_{10} + M_{20}$$

$\bar{M}_1$ and $\bar{M}_2$ — the average sizes of the first-stage units in the first and the second strata, respectively;

P_{1i} and P_{2j} — the selection probabilities at the first draw for the i-th first-stage unit in the first stratum and the j-th first-stage unit in the second stratum, respectively;

and

m_{1i} and m_{2j} — the numbers of second-stage units to be included in the sample from the selected first-stage units in the first and the second strata, respectively.

For convenience, however, we shall suppose that the first-stage unit selected from the first stratum is the c-th and that from the second the d-th unit.

Since $P_{1i} = M_{1i}/M_{10}$, clearly $\bar{y}_{m_{1c}}$ will represent an unbiased estimate of $\bar{y}_{1..}$, the population mean per second-stage unit of the first stratum; and similarly, since $P_{2j} = M_{2j}/M_{20}$, $\bar{y}_{m_{2d}}$ will represent an unbiased estimate of $\bar{y}_{2..}$ for the second stratum.

An unbiased estimate of the population mean $\bar{y}_{..}$ for the two strata together will be given by the weighted mean of $\bar{y}_{m_{1c}}$ and $\bar{y}_{m_{2d}}$ and may be denoted by $\bar{y}_w$, given by

$$\bar{y}_w = \lambda \bar{y}_{m_{1c}} + (1 - \lambda)\, \bar{y}_{m_{2d}} \tag{130}$$

where

$$\lambda = \frac{M_{10}}{M_0} \tag{131}$$

The variance of $\bar{y}_w$ will be

$$V(\bar{y}_w) = \lambda^2 V(\bar{y}_{m_{1c}}) + (1-\lambda)^2 V(\bar{y}_{m_{2d}})$$

$$= \lambda^2 \left\{\sigma_{1b}^2 + \sum_{i=1}^{N_1} P_{1i} \left(\frac{1}{m_{1i}} - \frac{1}{M_{1i}}\right) S_{1i}^2\right\}$$

$$+ (1-\lambda)^2 \left\{\sigma_{2b}^2 + \sum_{j=1}^{N_2} P_{2j} \left(\frac{1}{m_{2j}} - \frac{1}{M_{2j}}\right) S_{2j}^2\right\} \qquad (132)$$

where

$$\left.\begin{aligned} \sigma_{1b}^2 &= \sum_{i=1}^{N_1} P_{1i} (\bar{y}_{1i.} - \bar{y}_{1..})^2 \\ \sigma_{2b}^2 &= \sum_{j=1}^{N_2} P_{2j} (\bar{y}_{2j.} - \bar{y}_{2..})^2 \end{aligned}\right\} \qquad (133)$$

and

$$\left.\begin{aligned} S_{1i}^2 &= \frac{1}{M_{1i}-1} \sum_{r=1}^{M_{1i}} (y_{1ir} - \bar{y}_{1i.})^2 \\ S_{2j}^2 &= \frac{1}{M_{2j}-1} \sum_{r=1}^{M_{2j}} (y_{2jr} - \bar{y}_{2j.})^2 \end{aligned}\right\} \qquad (134)$$

Now, if the sample of two first-stage units were selected as an unstratified sample with probability proportional to the measure of size of the units, then an appropriate estimate of the population mean would be given by the simple arithmetic mean of cluster means in the sample, namely $\bar{y}_s$, and its variance by

$$V(\bar{y}_s) = \frac{1}{2} \left\{\sigma_b^2 + \sum_{l=1}^{N} P_l \left(\frac{1}{m_l} - \frac{1}{M_l}\right) S_l^2 \right\} \qquad (135)$$

where

$$\sigma_b^2 = \sum_{l=1}^{N} P_l (\bar{y}_{l.} - \bar{y}_{..})^2 \qquad (136)$$

and

$$S_l^2 = \frac{1}{M_l - 1} \sum_{r=1}^{M_l} (y_{lr} - \bar{y}_{l.})^2 \qquad (137)$$

It follows from the relation $P_{1i} = M_{1i}/M_{10}$ and (131) that

$$P_i = \lambda P_{1i} \qquad (i = 1, 2, \ldots, N_1) \qquad (138)$$

for units in the first stratum, and

$$P_i = (1 - \lambda)\, P_{2j} \qquad (j = 1, 2, \ldots, N_2) \qquad (139)$$

for units in the second stratum.

Now equation (120) of the preceding section shows that when $P_{t.} = \lambda_t$ and $N_1 = N_2$ and the allocation of first-stage units to the two strata is in proportion to $\lambda : (1 - \lambda)$, (135) provides an upper bound to the actual variance of the estimate of a stratified sample. For evaluating it from the selected sample, we require the estimate of σ_b^2. We write

$$\sigma_b^2 = \sum_{i=1}^{N} P_i\,(\bar{y}_{i.} - \bar{y}_{..})^2$$

$$= \sum_{i=1}^{N_1} \lambda P_{1i}\,(\bar{y}_{1i.} - \bar{y}_{1..} + \bar{y}_{1..} - \bar{y}_{..})^2$$

$$+ \sum_{j=1}^{N_2} (1 - \lambda)\, P_{2j}\,(\bar{y}_{2j.} - \bar{y}_{2..} + \bar{y}_{2..} - \bar{y}_{..})^2$$

$$= \sum_{i=1}^{N_1} \lambda P_{1i}\,\{(\bar{y}_{1i.} - \bar{y}_{1..})^2 + (\bar{y}_{1..} - \bar{y}_{..})^2\}$$

$$+ \sum_{j=1}^{N_2} (1 - \lambda)\, P_{2j}\,\{(\bar{y}_{2j.} - \bar{y}_{2..})^2 + (\bar{y}_{2..} - \bar{y}_{..})^2\}$$

$$= \lambda \sigma_{1b}^2 + (1 - \lambda)\, \sigma_{2b}^2 + \lambda\,(\bar{y}_{1..} - \bar{y}_{..})^2 + (1 - \lambda)\,(\bar{y}_{2..} - \bar{y}_{..})^2$$

$$= \lambda \sigma_{1b}^2 + (1 - \lambda)\, \sigma_{2b}^2 + \lambda \bar{y}_{1..}^2 + (1 - \lambda)\, \bar{y}_{2..}^2 - \bar{y}_{..}^2 \qquad (140)$$

We know that

$$V(\bar{y}_{m_{1c}}) = E(\bar{y}_{m_{1c}}^2) - \bar{y}_{1..}^2$$

$$= \sigma_{1b}^2 + \sum_{i=1}^{N_1} P_{1i}\left(\frac{1}{m_{1i}} - \frac{1}{M_{1i}}\right) S_{1i}^2 \qquad (141)$$

and

$$V(\bar{y}_{m_{2d}}) = E(\bar{y}_{m_{2d}}^2) - \bar{y}_{2..}^2$$

$$= \sigma_{2b}^2 + \sum_{j=1}^{N_2} P_{2j} \left(\frac{1}{m_{2j}} - \frac{1}{M_{2j}}\right) S_{2j}^2 \tag{142}$$

giving us

$$\text{Est. } \bar{y}_{1..}^2 = \bar{y}_{m_{1c}}^2 - \hat{\sigma}_{1b}^2 - \left(\frac{1}{m_{1c}} - \frac{1}{M_{1c}}\right) \hat{S}_{1c}^2 \tag{143}$$

and

$$\text{Est. } \bar{y}_{2..}^2 = \bar{y}_{m_{2d}}^2 - \hat{\sigma}_{2b}^2 - \left(\frac{1}{m_{2d}} - \frac{1}{M_{2d}}\right) \hat{S}_{2d}^2 \tag{144}$$

Also, from (132), we get

$$\text{Est. } \bar{y}_{..}^2 = \bar{y}_{w}^2 - \lambda^2 \left\{ \hat{\sigma}_{1b}^2 + \left(\frac{1}{m_{1c}} - \frac{1}{M_{1c}}\right) \hat{S}_{1c}^2 \right\}$$

$$- (1 - \lambda)^2 \left\{ \hat{\sigma}_{2b}^2 + \left(\frac{1}{m_{2d}} - \frac{1}{M_{2d}}\right) \hat{S}_{2d}^2 \right\} \tag{145}$$

Substituting for $\bar{y}_{1..}^2$, $\bar{y}_{2..}^2$ and $\bar{y}_{..}^2$ from (143), (144) and (145) in (140), we get on simplification

$$\text{Est. } \sigma_b^2 = \lambda \bar{y}_{m_{1c}}^2 + (1 - \lambda)\, \bar{y}_{m_{2d}}^2 - \bar{y}_{w}^2 + \lambda^2 \hat{\sigma}_{1b}^2 + (1 - \lambda)^2\, \hat{\sigma}_{2b}^2$$

$$- \lambda (1 - \lambda) \left(\frac{1}{m_{1c}} - \frac{1}{M_{1c}}\right) \hat{S}_{1c}^2$$

$$- \lambda (1 - \lambda) \left(\frac{1}{m_{2d}} - \frac{1}{M_{2d}}\right) \hat{S}_{2d}^2 \tag{146}$$

Estimates of σ_{1b}^2 and σ_{2b}^2 cannot be calculated from a sample of one each. In a reasonably efficient scheme of stratification, however, σ_{1b}^2 and σ_{2b}^2 can each be expected to be smaller than σ_b^2. Replacing them by their upper bound, namely σ_b^2, we then obtain for the estimate of σ_b^2 an upper bound given by

$$\overline{\text{Est. } \sigma_b^2} = \frac{1}{2(1 - \lambda)} (\bar{y}_{m_{1c}} - \bar{y}_{w})^2 + \frac{1}{2\lambda} (\bar{y}_{m_{2d}} - \bar{y}_{w})^2$$

$$- \frac{1}{2} \left\{ \left(\frac{1}{m_{1c}} - \frac{1}{M_{1c}}\right) \hat{S}_{1c}^2 + \left(\frac{1}{m_{2d}} - \frac{1}{M_{2d}}\right) \hat{S}_{2d}^2 \right\} \tag{147}$$

Now, an estimate of the variance of the sample mean, if the sample were selected as an unstratified sample, is obtained from (135), being given by

$$\text{Est. } V(\bar{y}_s) = \frac{1}{2}\left[\hat{\sigma}_b^2 + \lambda\left(\frac{1}{m_{1c}} - \frac{1}{M_{1c}}\right)\hat{S}_{1c}^2 + (1-\lambda)\left(\frac{1}{m_{2d}} - \frac{1}{M_{2d}}\right)\hat{S}_{2d}^2\right] \tag{148}$$

Substituting for $\hat{\sigma}_b^2$ from (147) and replacing $\hat{S}_{1c}^2$ and $\hat{S}_{2d}^2$ by s_{1c}^2 and s_{2d}^2 respectively, we get

$$\overline{\text{Est. } V(\bar{y}_s)} = \frac{1}{2}\left[\frac{1}{2(1-\lambda)}(\bar{y}_{m_{1c}} - \bar{y}_w)^2 + \frac{1}{2\lambda}(\bar{y}_{m_{2d}} - \bar{y}_w)^2 - (\tfrac{1}{2} - \lambda)\left(\frac{1}{m_{1c}} - \frac{1}{M_{1c}}\right)s_{1c}^2 + (\tfrac{1}{2} - \lambda)\left(\frac{1}{m_{2d}} - \frac{1}{M_{2d}}\right)s_{2d}^2\right] \tag{149}$$

For $\lambda = \frac{1}{2}$, we have the inequality

$$\sigma_b^2 \geqslant \tfrac{1}{2}(\sigma_{1b}^2 + \sigma_{2b}^2)$$

Consequently (147) will always provide an upper bound to the estimate of σ_b^2 and hence (149) will provide an upper bound to the actual error variance of the mean of a stratified sample. Substituting $\lambda = \frac{1}{2}$ in (149), we get the simple expression

$$\overline{\text{Est. } V(\bar{y}_s)} = \tfrac{1}{4}(\bar{y}_{m_{1c}} - \bar{y}_{m_{2d}})^2 \tag{150}$$

It should be emphasized that the expression (150) is only an upper bound and does not necessarily provide a satisfactory approximation to the error variance. In fact, in most surveys where stratification is effective, it will be found to result in a considerable over-estimate of the actual error.

8.13 Sub-Sampling with Varying Probabilities of Selection at Each Stage

Lastly, we shall consider a sub-sampling system in which units at each stage of sampling are selected with replacement, with probabilities proportional to measures of their sizes. It is easily shown that under this system each unit measure of size gets an equal chance of being included in the sample and in consequence, a simple arithmetic mean of the ratios of the observed value y to the measure of size x for the units in the sample provides an unbiased estimate of the population ratio of the total of y to the total of x.

Let

x_{ij} denote the measure of size of the j-th second-stage unit within the i-th first-stage unit;

X_i the measure of size of the i-th first-stage unit, so that

$$X_i = \sum_{j=1}^{M_i} x_{ij} \quad \text{and} \quad X = \sum_{i=1}^{N} X_i$$

Y_i the total of y for the i-th first-stage unit, so that

$$Y_i = \sum_{j=1}^{M_i} y_{ij} \quad \text{and} \quad Y = \sum_{i=1}^{N} Y_i$$

r_{ij} the ratio of y to x for the j-th second-stage unit within the i-th first-stage unit given by

$$r_{ij} = \frac{y_{ij}}{x_{ij}}$$

R_i the ratio of the total of y to the total of x for the i-th first-stage unit given by

$$R_i = \frac{Y_i}{X_i}$$

and

R the ratio of the population total of y to the population total of x given by

$$R = \frac{Y}{X}$$

It is easily seen that

$$E(r_{ij} \mid i) = R_i \tag{151}$$

and

$$E(R_i) = R \tag{152}$$

For,

$$\begin{aligned} E(r_{ij}) &= \sum_{j=1}^{M_i} \frac{x_{ij}}{X_i}\, r_{ij} \\ &= \frac{1}{X_i} \sum_{j=1}^{M_i} y_{ij} \\ &= \frac{Y_i}{X_i} \\ &= R_i \end{aligned}$$

and

$$\begin{aligned} E(R_i) &= \sum_{i=1}^{N} \frac{X_i}{X}\, \frac{Y_i}{X_i} \\ &= \frac{Y}{X} \\ &= R \end{aligned}$$

Consider now the simple arithmetic mean of the ratios r_{ij} given by

$$\bar{r}_{nm} = \frac{1}{nm} \sum_{i}^{n} \sum_{j}^{m} \frac{y_{ij}}{x_{ij}} \tag{153}$$

Clearly then, using (151) and (152), we get

$$\begin{aligned} E(\bar{r}_{nm}) &= \frac{1}{nm}\, E\left\{\sum^{n} E\left(\sum^{m} r_{ij} \mid i\right)\right\} \\ &= \frac{1}{nm}\, E\left\{m \sum^{n} R_i\right\} \\ &= \frac{1}{n} \sum^{n} E(R_i) \\ &= R \end{aligned} \tag{154}$$

To obtain the sampling variance of $\bar{r}_{nm}$, we write

$$V(\bar{r}_{nm}) = E(\bar{r}_{nm} - R)^2$$

$$= E(\bar{r}_{nm} - \bar{R}_n + \bar{R}_n - R)^2$$

where $\bar{R}_n = (1/n) \overset{n}{\Sigma} R_i$,

$$= E\{(\bar{r}_{nm} - \bar{R}_n)^2 + (\bar{R}_n - R)^2 + 2(\bar{r}_{nm} - \bar{R}_n)(\bar{R}_n - R)\}$$

$$= E(\bar{r}_{nm} - \bar{R}_n)^2 + E(\bar{R}_n - R)^2 \tag{155}$$

since for a fixed sample of n first-stage units $E(\bar{r}_{nm}) = \bar{R}_n$ and in consequence the last term is zero.

Taking the first term in (155), we have

$$E(\bar{r}_{nm} - \bar{R}_n)^2 = E\left\{\frac{1}{n}\sum^{n}(\bar{r}_{im} - R_i)\right\}^2$$

$$= \frac{1}{n^2}E\left\{\sum^{n}(\bar{r}_{im} - R_i)^2 + \sum_{i \neq i'}^{n}(\bar{r}_{im} - R_i)(\bar{r}_{i'm} - R_{i'})\right\}$$

$$= \frac{1}{n^2}E\left\{\sum^{n}(\bar{r}_{im} - R_i)^2\right\} \tag{156}$$

the second term vanishing since sampling within the i-th and i'-th units is carried out independently.

From (87) of Chapter VI, we may write

$$E\{(\bar{r}_{im} - R_i)^2 \mid i\} = \frac{\sigma_i^2}{m} \tag{157}$$

where

$$\sigma_i^2 = \sum_{j=1}^{M_i} \frac{x_{ij}}{X_i}(r_{ij} - R_i)^2 \tag{158}$$

sampling units with equal probability and using the simple arithmetic mean to estimate the yield rate.

8.14* Sampling without Replacement at Each Stage

So far we have assumed that the first-stage units are selected with replacement. Clearly, the efficiency of a sampling procedure is reduced by including the same unit twice or oftener in the sample. One method of improving the efficiency is to group together into strata first-stage units of about the same size. However, as pointed out in Section 7.15, stratification by size of unit may not be always feasible and even where such stratification is attempted the units within strata may still show considerable variation in size. In this situation sampling without replacement within strata with probability proportional to the measure of the size of the units can be used with considerable gains in efficiency. In this section we shall extend the theory to the case when the first-stage units are selected without replacement with varying probabilities of selection and from each selected unit a simple random sample of predetermined size is drawn without replacement.

Let

$$z_{ij} = \frac{nM_i y_{ij}}{M_0 E(\alpha_i)} \tag{169}$$

where $E(\alpha_i)$, as defined in Section $2a.4$, denotes the probability of including the i-th first-stage unit in a sample of n. Consider the estimate

$$\bar{z}_s = \frac{1}{n} \sum^{n} \bar{z}_{i(m_i)} \tag{170}$$

Clearly, $\bar{z}_s$ gives an unbiased estimate of $\bar{y}_{..}$. For,

$$E(\bar{z}_s) = E\left\{\frac{1}{n} \sum^{n} \bar{z}_{i(m_i)}\right\}$$

$$= E\left\{\frac{1}{n} \sum^{n} E(\bar{z}_{i(m_i)} \mid i)\right\}$$

$$= E\left\{\frac{1}{n}\sum^{n}\bar{z}_{i.}\right\}$$

$$= \frac{1}{n}\sum_{i=1}^{N} E(\alpha_i)\,\bar{z}_{i.} \tag{171}$$

Substituting for $\bar{z}_{i.}$ from (169), we have

$$E(\bar{z}_s) = \bar{y}_{..} \tag{172}$$

To obtain the variance, we write

$$V(\bar{z}_s) = E(\bar{z}_s^2) - \bar{y}_{..}^2$$

$$= E\left\{\left(\frac{1}{n}\sum^{n}\bar{z}_{i(m_i)}\right)^2\right\} - \bar{y}_{..}^2$$

$$= E\left\{\left(\sum^{n}\frac{M_i}{M_0}\,\frac{\bar{y}_{i(m_i)}}{E(\alpha_i)}\right)^2\right\} - \bar{y}_{..}^2$$

$$= E\left[\sum^{n}\frac{M_i^2\bar{y}^2_{i(m_i)}}{M_0^2\{E(\alpha_i)\}^2} + \sum_{i\neq j}^{n}\frac{M_iM_j\bar{y}_{i(m_i)}\bar{y}_{j(m_j)}}{M_0^2E(\alpha_i)\,E(\alpha_j)}\right] - \bar{y}_{..}^2$$

$$= E\left[\sum^{n}\frac{M_i^2E\{\bar{y}^2_{i(m_i)} \mid i\}}{M_0^2\{E(\alpha_i)\}^2}\right.$$

$$\left. + \sum_{i\neq j}^{n}\frac{M_iM_jE\{\bar{y}_{i(m_i)}\bar{y}_{j(m_j)} \mid i, j\}}{M_0^2E(\alpha_i)\,E(\alpha_j)}\right] - \bar{y}_{..}^2$$

$$= E\left[\sum^{n}\frac{M_i^2}{M_0^2\{E(\alpha_i)\}^2}\left\{\bar{y}_{i.}^2 + \left(\frac{1}{m_i} - \frac{1}{M_i}\right)S_i^2\right\}\right]$$

$$+ E\left[\sum_{i\neq j}^{n}\frac{M_iM_j\bar{y}_{i.}\bar{y}_{j.}}{M_0^2E(\alpha_i)\,E(\alpha_j)}\right] - \bar{y}_{..}^2$$

$$= \sum_{i=1}^{N} \frac{M_i^2}{M_0^2\, E(a_i)} \left\{\bar{y}_{i.}^2 + \left(\frac{1}{m_i} - \frac{1}{M_i}\right) S_i^2\right\} + \sum_{i\neq j=1}^{N} \frac{E(a_ia_j)}{E(a_i)\, E(a_j)} \cdot \frac{M_iM_j\bar{y}_{i.}\bar{y}_{j.}}{M_0^2} - \bar{y}_{..}^2 \tag{173}$$

where $E(a_ia_j)$, as defined in Section $2a.5$, denotes the probability of including the i-th and the j-th units in the sample.

Expanding $\bar{y}_{..}^2$, we may rewrite (173) as

$$V(\bar{z}_s) = \sum_{i=1}^{N} \frac{1 - E(a_i)}{E(a_i)} \frac{M_i^2\bar{y}_{i.}^2}{M_0^2} + \sum_{i\neq j=1}^{N} \frac{E(a_ia_j) - E(a_i)\, E(a_j)}{E(a_i)\, E(a_j)} \frac{M_iM_j\bar{y}_{i.}\bar{y}_{j.}}{M_0^2} + \sum_{i=1}^{N} \frac{M_i^2}{M_0^2} \left(\frac{1}{m_i} - \frac{1}{M_i}\right) \frac{S_i^2}{E(a_i)} \tag{174}$$

This expression was first given by Horvitz and Thompson (1952).

For the case of simple random sampling, we know from (19) and (31) of Chapter II that

$$E(a_i) = \frac{n}{N} \quad \text{and} \quad E(a_ia_j) = \frac{n(n-1)}{N(N-1)}$$

Substituting in (170), we notice that the estimate $\bar{z}_s$ reduces to $\bar{y}_s'$ of Section 7.11, namely,

$$\bar{z}_s = \bar{y}_s' = \frac{1}{n\bar{M}} \sum^{n} M_i\bar{y}_{i(m_i)} \tag{175}$$

and the expression for the variance becomes identical with that given by (91) of Section 7.12, as expected. Thus,

$$V(\bar{z}_s) = \frac{1}{N}\left(\frac{1}{n} - \frac{1}{N}\right) \sum_{i=1}^{N} \frac{M_i^2 \bar{y}_{i.}^2}{\bar{M}^2}$$

$$- \frac{1}{N}\left(\frac{1}{n} - \frac{1}{N}\right) \frac{1}{N-1} \sum_{i \neq j=1}^{N} \frac{M_i M_j \bar{y}_{i.} \bar{y}_{j.}}{\bar{M}^2}$$

$$+ \frac{1}{nN} \sum_{i=1}^{N} \frac{M_i^2}{\bar{M}^2} \left(\frac{1}{m_i} - \frac{1}{M_i}\right) \mathbf{S}_i^2$$

$$= \left(\frac{1}{n} - \frac{1}{N}\right) \mathbf{S}_b'^2 + \frac{1}{nN} \sum_{i=1}^{N} \frac{M_i^2}{\bar{M}^2} \left(\frac{1}{m_i} - \frac{1}{M_i}\right) \mathbf{S}_i^2 \quad (176)$$

where

$$\mathbf{S}_b'^2 = \frac{1}{N-1} \sum_{i=1}^{N} \left(\frac{M_i}{\bar{M}} \bar{y}_{i.} - \bar{y}_{..}\right)^2 \quad (177)$$

In developing these formulæ, we assumed that m_i, the number to be sampled from the i-th first-stage unit, is known in advance. In order to determine it we use the principle of minimizing the cost of the survey for a given precision, or alternatively, of maximizing the precision for a given cost. The application of this principle when the cost is represented by (20) is straightforward and follows step by step the analysis shown in Section 8.4. It will be found that the optimum value of m_i is approximately given by

$$m_i \frac{E(a_i)}{M_i} = \text{constant} \quad (178)$$

When $E(a_i)$ is proportional to M_i, as will usually be our attempt to make it, m_i will be constant irrespective of the first-stage unit included in the sample. One important simplification arising from this design is that the estimate $\bar{z}_S$ reduces to the simple arithmetic mean of the nm y-values in the sample.

An unbiased estimate of $V(\bar{z}_S)$ is easily computed. We write, from (174),

$$\begin{aligned}
\text{Est. } V(\bar{z}_s) &= \sum^{n} \frac{1 - E(a_i)}{\{E(a_i)\}^2} \frac{M_i^2}{M_0^2} \text{ Est. } (\bar{y}_{i.}^2 \mid i) \\
&\quad + \sum_{i \neq j}^{n} \frac{E(a_i a_j) - E(a_i)\, E(a_j)}{E(a_i a_j)\, E(a_i)\, E(a_j)} \frac{M_i M_j}{M_0^2} \\
&\qquad \times \text{ Est. } (\bar{y}_{i.}\bar{y}_{j.} \mid i, j) \\
&\quad + \sum^{n} \frac{1}{\{E(a_i)\}^2} \frac{M_i^2}{M_0^2} \left(\frac{1}{m_i} - \frac{1}{M_i}\right) \text{ Est. } (S_i^2 \mid i) \\
&= \sum^{n} \frac{1 - E(a_i)}{\{E(a_i)\}^2} \frac{M_i^2}{M_0^2} \left\{\bar{y}^2_{i(mi)} - \left(\frac{1}{m_i} - \frac{1}{M_i}\right) s_i^2\right\} \\
&\quad + \sum_{i \neq j}^{n} \left\{\frac{1}{E(a_i)\, E(a_j)} - \frac{1}{E(a_i a_j)}\right\} \frac{M_i M_j}{M_0^2} \\
&\qquad \times \bar{y}_{i(mi)}\bar{y}_{j(mj)} \\
&\quad + \sum^{n} \frac{1}{\{E(a_i)\}^2} \frac{M_i^2}{M_0^2} \left(\frac{1}{m_i} - \frac{1}{M_i}\right) s_i^2 \\
&= \sum^{n} \frac{1 - E(a_i)}{\{E(a_i)\}^2} \frac{M_i^2 \bar{y}^2_{i(mi)}}{M_0^2} \\
&\quad + \sum_{i \neq j}^{n} \left\{\frac{1}{E(a_i)\, E(a_j)} - \frac{1}{E(a_i a_j)}\right\} \\
&\quad \times \frac{M_i M_j}{M_0^2} \bar{y}_{i(mi)}\bar{y}_{j(mj)} \\
&\quad + \sum^{n} \frac{1}{E(a_i)} \frac{M_i^2}{M_0^2} \left(\frac{1}{m_i} - \frac{1}{M_i}\right) s_i^2 \qquad (179)
\end{aligned}$$

An alternative estimate of the variance based on a linear combination of the squared differences in the sample, which appears to be better than the one given above, can be formed following the device mentioned in the foot-note on page 71 of Chapter II. Thus, we rewrite (174) as

$$V(\bar{z}_s) = \frac{1}{2n^2}\sum_{i\neq j=1}^{N}\left\{E(a_i)\,E(a_j) - E(a_ia_j)\right\}(\bar{z}_{i.} - \bar{z}_{j.})^2$$

$$+ \frac{1}{n^2}\sum_{i=1}^{N} E(a_i)\left(\frac{1}{m_i} - \frac{1}{M_i}\right) S_{iz}^2$$

Hence

$$\text{Est. } V(\bar{z}_s) = \frac{1}{2n^2}\sum_{i\neq j}^{n}\left\{\frac{E(a_i)\,E(a_j) - E(a_ia_j)}{E(a_ia_j)}\right\}$$

$$\times \text{ Est. }\{(\bar{z}_{i.} - \bar{z}_{j.})^2 \mid i, j\}$$

$$+ \frac{1}{n^2}\sum^{n}\left(\frac{1}{m_i} - \frac{1}{M_i}\right) s_{iz}^2$$

$$= \frac{1}{2n^2}\sum_{i\neq j}^{n}\left\{\frac{E(a_i)\,E(a_j) - E(a_ia_j)}{E(a_ia_j)}\right\}$$

$$\times \left\{(\bar{z}_{i(m_i)} - \bar{z}_{j(m_j)})^2 - \left(\frac{1}{m_i} - \frac{1}{M_i}\right) s_{iz}^2\right.$$

$$\left. - \left(\frac{1}{m_j} - \frac{1}{M_j}\right) s_{jz}^2\right\}$$

$$+ \frac{1}{n^2}\sum^{n}\left(\frac{1}{m_i} - \frac{1}{M_i}\right) s_{iz}^2 \tag{180}$$

The theory of sampling without replacement developed in this section is difficult to use in practice on account of the heavy computations involved in evaluating $E(a_i)$ and $E(a_ia_j)$. For the important case of samples of two within each stratum, explicit expressions for $E(a_i)$ and $(E\,a_ia_j)$ in terms of the selection probabilities $P_1, P_2, \ldots, P_N$ have been given in Section 2b.4 and are relatively easy to compute. For larger samples the use of the estimate appropriate for sampling with replacement, introducing the usual finite multiplier for calculating the error variance, is probably sufficiently satisfactory (Yates, 1949).

REFERENCES

1. Hansen, M. H. and Hurwitz, W. N. (1943) "On the Theory of Sampling from Finite Populations," *Ann. Math. Statist.*, **14**, 333–62.

2. ——— (1949) .. "On the Determination of Optimum Probabilities in Sampling," *Ann. Math. Statist.*, **20**, 426–32.

3. Sukhatme, P. V. and Narain, R. D. (1952) "Sampling with Replacement," *Jour. Ind. Soc. Agr. Statist.*, **4**, 42–49.

4. Mahalanobis, P. C. (1945) "Report on the Bihar Crop Survey, 1943–44," *Sankhya*, **7**, 29–106.

5. ——— (1948) .. "Report on Bengal Crop Survey, 1944–45," *Department of Agriculture, Forests and Fisheries, West Bengal.*

6. I.C.A.R., New Delhi (1950) "Report on the Sample Survey for Estimation of Acreage under Principal Crops in Orissa" (*Unpublished*).

7. Horvitz, D. G. and Thompson, D. J. (1952) "A Generalisation of Sampling without Replacement from a Finite Universe," *Jour. Amer. Statist. Assoc.*, **47**, 663–85.

8. Yates, F. (1949) .. *Sampling Methods for Censuses and Surveys*, Charles Griffin & Co., Ltd., London.

CHAPTER IX

SYSTEMATIC SAMPLING

9.1 Introduction

So far we have considered methods of sampling in which the successive units (whether elements or clusters) were selected with the help of random numbers. We shall now consider a method of sampling in which only the first unit is selected with the help of random numbers, the rest being selected automatically according to a predetermined pattern. The method is known as *systematic sampling*.

The pattern usually followed in selecting a systematic sample is a simple pattern involving regular spacing of units. Thus, suppose a population consists of N units, serially numbered from 1 to N. Suppose further that N is expressible as a product of two integers k and n, so that $N = kn$. Draw a random number less than k, say i, and select the unit with the corresponding serial number and every k-th unit in the population thereafter. Clearly, the sample will contain the n units i, $i + k$, $i + 2k$, . . ., $i + \overline{n - 1}k$, and is known as a *systematic sample*. The selection of every k-th strip in forest sampling for the estimation of timber, the selection of a corn field, every k-th mile apart, for observation on incidence of borers, the selection of every k-th time-interval for observing the number of fishing craft landing on the coast, the selection of every k-th punched card for advance tabulation or of every k-th village from a list of villages, after the first unit is chosen with the help of random numbers less than k, are all examples of systematic sampling. In the first three examples, the sequence of numbering is determined by Nature, the first two providing examples of distribution in space while the third that of distribution in time. In the fourth and the fifth, the ordering may be either alphabetical or arbitrary approximating to a random distribution. In the latter case, a systematic sample will obviously be equivalent to a random sample. The method is extensively used in practice on account of its low cost and simplicity in the selection of the sample. The latter consideration is particularly important in situations where the selection of a sample is carried out by the field staff themselves. A systematic sample also offers great advantages in organizing control over field work.

In a systematic sample, as noted already, the relative position in the population of the different units included in the sample is fixed. There is consequently no risk in the method that any large contiguous part of the population will fail to be represented. Indeed, the method will give an evenly spaced sample and is, therefore, likely to give a more precise estimate of the population mean than a random sample unless the k-th units constituting the sample happen to be alike or correlated. The method resembles stratified sampling in that one sampling unit is selected from each stratum of k consecutive units. In reality, however, the resemblance is only casual. In stratified sampling the unit to be selected from each stratum is randomly drawn, in systematic sampling its position relative to the unit in the first stratum is predetermined. Unless, therefore, the units in each stratum are randomly listed, a systematic sample will not be equivalent to a stratified random sample.

Systematic sampling strictly resembles cluster sampling, a systematic sample being equivalent to a sample of one cluster selected out of the k clusters of n units each, shown in the schematic diagram below in the form of k columns of n each:

Schematic Diagram Showing the Serial Number of the Unit in the Population in Terms of the Cluster Number and its Serial Position in the Cluster

	Cluster Number i	1	2	…	i	…	k
Row No.	1	1	2	…	i	…	k
	2	$1+k$	$2+k$	…	$i+k$	…	$2k$
	3	$1+2k$	$2+2k$	…	$i+2k$	…	$3k$
	.	.	.	…	.	…	.
	.	.	.	…	.	…	.
	.	.	.	…	.	…	.
	j	$1+(j-1)k$	$2+(j-1)k$	…	$i+(j-1)k$	…	jk
	.	.	.	…	.	…	.
	.	.	.	…	.	…	.
	.	.	.	…	.	…	.
	n	$1+(n-1)k$	$2+(n-1)k$	…	$i+(n-1)k$	…	nk

(1)

Since the first number less than or equal to k is to be chosen at random, every one of the k columns gets an equal chance of being chosen as the systematic sample. It follows that the theory of systematic sampling can be deduced from the theory of cluster sampling dealt with in Chapter VI.

In presenting the theory in this chapter, we shall assume that $N = nk$, where n is the size of the sample and k is an integer. In practice N may not be so expressible, and the results presented in this chapter may not be strictly applicable. However, the disturbance is not likely to be large unless n is small.

9.2 The Sample Mean and its Variance

Let

y_{ij} denote the observation on the unit bearing the serial number $i + (j - 1)k$ in the population ($i = 1, 2, \ldots, k$; $j = 1, 2, \ldots, n$);

$\bar{y}_{i.}$ the sample mean

$$= \frac{1}{n} \sum_{j=1}^{n} y_{ij} \tag{2}$$

and

$\bar{y}_{..}$ the population mean

$$= \frac{1}{nk} \sum_{i=1}^{k} \sum_{j=1}^{n} y_{ij} \tag{3}$$

Since the probability of selecting the i-th column as the systematic sample is $1/k$, it follows that

$$\begin{aligned} E(\bar{y}_{i.}) &= \frac{1}{k} \sum_{i=1}^{k} \bar{y}_{i.} \\ &= \bar{y}_{..} \end{aligned} \tag{4}$$

showing that a systematic sample provides an unbiased estimate of the population mean.

The variance is given by

$$V(\bar{y}_{i.})_{Sy} = E\{(\bar{y}_{i.} - \bar{y}_{..})^2\}$$

$$= \frac{1}{k} \sum_{i=1}^{k} (\bar{y}_{i.} - \bar{y}_{..})^2 \tag{5}$$

which can also be written as

$$= \frac{k-1}{k} S_c^2 \tag{6}$$

where S_c^2 denotes the mean square between column means, the letter c standing for a column.

9.3 Comparison of Systematic with Random Sampling

The variance of the mean of a random sample of n units chosen from a population of size N is known to be given by

$$V(\bar{y}_n)_R = \left(\frac{1}{n} - \frac{1}{N}\right) S^2 \tag{7}$$

where S^2 is the mean square between units in the population. This is not directly comparable with (6), and it is therefore important to express the variance of a systematic sample in an alternative form suitable for this comparison. We write

$$V(\bar{y}_{i.})_{Sy} = \frac{1}{k} \sum_{i=1}^{k} (\bar{y}_{i.} - \bar{y}_{..})^2$$

$$= \frac{1}{k} \sum_{i=1}^{k} \left\{\frac{1}{n} \sum_{j=1}^{n} y_{ij} - \bar{y}_{..}\right\}^2$$

$$= \frac{1}{kn^2} \sum_{i=1}^{k} \left\{\sum_{j=1}^{n} (y_{ij} - \bar{y}_{..})\right\}^2$$

$$= \frac{1}{kn^2} \sum_{i=1}^{k} \left\{ \sum_{j=1}^{n} (y_{ij} - \bar{y}_{..})^2 + \sum_{j \neq j'=1}^{n} (y_{ij} - \bar{y}_{..})(y_{ij'} - \bar{y}_{..}) \right\}$$

$$= \frac{1}{kn^2} \left\{ (nk-1) S^2 + \sum_{i=1}^{k} \sum_{j \neq j'=1}^{n} (y_{ij} - \bar{y}_{..})(y_{ij'} - \bar{y}_{..}) \right\} \qquad (8)$$

Now, by definition, the intra-class correlation between units of a column is given by

$$\rho = \frac{E(y_{ij} - \bar{y}_{..})(y_{ij'} - \bar{y}_{..})}{E(y_{ij} - \bar{y}_{..})^2}$$

$$= \frac{\sum\limits_{i=1}^{k} \sum\limits_{j \neq j'=1}^{n} (y_{ij} - \bar{y}_{..})(y_{ij'} - \bar{y}_{..})}{kn(n-1)} \cdot \frac{kn}{(kn-1) S^2} \qquad (9)$$

or

$$\sum_{i=1}^{k} \sum_{j \neq j'=1}^{n} (y_{ij} - \bar{y}_{..})(y_{ij'} - \bar{y}_{..}) = (n-1)(kn-1)\rho S^2 \qquad (10)$$

Substituting from (10) in (8), we get

$$V(\bar{y}_{i.})_{Sy} = \frac{kn-1}{kn} \cdot \frac{S^2}{n} \{1 + \rho(n-1)\} \qquad (11)$$

which is a convenient form for purposes of comparison with the variance of the mean of a random sample. The variance of a systematic sample relative to that of a random sample is seen to be

$$\frac{V_{Sy}}{V_R} = \frac{(nk-1)\{1 + \rho(n-1)\}}{n(k-1)} \qquad (12)$$

We notice that the relative precision depends on the value of ρ. For $\rho = -1/(kn-1)$, the two methods give estimates of equal precision; for ρ greater than $-1/(kn-1)$, systematic sampling is less accurate than random sampling; while for ρ less than $-1/(kn-1)$, systematic sampling is superior to random sampling. The minimum value which ρ can take is $-1/(n-1)$, when the variance of a systematic sample will be zero, and the reduction in variance over random sampling will therefore be 100%. The maximum value which ρ can assume is 1, when the efficiency of systematic relative to random sampling will be given by $(k-1)/(nk-1)$.

In general, however, it is difficult to know what values ρ will take in populations distributed in space or time, and no general conclusions can therefore be drawn about the relative efficiency of systematic and random sampling. On the other hand, for populations for which the lists of units are prepared in alphabetical or arbitrary order and where there is little likelihood of the lists corresponding to any physical distribution, we may assume the intra-class correlation to provide a good estimate of its average value in randomly formed columns, namely $-1/(nk-1)$, and hence expect systematic and random sampling to give results of about equal precision on an average.

It is instructive to express the variance of the systematic sample in terms of a further break-up of the intra-class correlation coefficient. It will be noticed from (10) that ρ is expressed as the sum of $kn\,(n-1)$ products of y deviations: $2k\,(n-1)$ of these products relate to y deviations separated by one row, $2k\,(n-2)$ of these products relate to y deviations separated by two rows, etc. We may, therefore, write (10) as

$$(n-1)\,(kn-1)\,\rho S^2 = 2\sum_{i=1}^{k}\sum_{a=1}^{n-1}\sum_{j=1}^{n-a}(y_{ij}-\bar{y}_{..})\,(y_{i\,j+a}-\bar{y}_{..})$$

$$= 2\sum_{a=1}^{n-1}k\,(n-a)\,\rho_a S^2\cdot\frac{kn-1}{kn}$$

where ρ_a is defined as

$$\rho_a = \frac{E\,(y_{ij}-\bar{y}_{..})\,(y_{i\,j+a}-\bar{y}_{..})}{E\,(y_{ij}-\bar{y}_{..})^2} \qquad (13)$$

and is called the non-circular serial correlation coefficient for lag $k\alpha$. Hence, we have

$$\rho = \frac{2}{n(n-1)} \sum_{\alpha=1}^{n-1} (n-\alpha)\,\rho_{\alpha} \tag{14}$$

Substituting for ρ from (14) in (11), we get

$$V(\bar{y}_{i.})_{sy} = \frac{kn-1}{kn} \cdot \frac{S^2}{n} \left\{1 + \frac{2}{n} \sum_{\alpha=1}^{n-1} (n-\alpha)\,\rho_{\alpha}\right\} \tag{15}$$

The expression is due to the Madows (1944).

9.4 Comparison of Systematic with Stratified Random Sampling

We shall consider the population as divided into n strata corresponding to the n rows of the schematic diagram (1), and suppose that one unit is randomly drawn from each one of these strata, thus giving us a stratified sample of n. Clearly, the variance of the mean of this sample will be

$$V(\bar{y}_{w})_{S} = \frac{1}{n}\left(1 - \frac{1}{k}\right) S_{wr}^{2} \tag{16}$$

where S_{wr}^2 is the mean square between units within rows, defined by

$$S_{wr}^{2} = \frac{1}{n(k-1)} \sum_{j=1}^{n} \sum_{i=1}^{k} (y_{ij} - \bar{y}_{.j})^2 \tag{17}$$

To examine how this compares with the variance of a systematic sample we shall first express the latter in a form suitable for direct comparison with (16).

Equation (5) can be written as

$$V(\bar{y}_{i.})_{sy} = \frac{1}{k} \sum_{i=1}^{k} \left\{\frac{1}{n} \sum_{j=1}^{n} y_{ij} - \frac{1}{n} \sum_{j=1}^{n} \bar{y}_{.j}\right\}^2$$

$$= \frac{1}{kn^2} \sum_{i=1}^{k} \left\{ \sum_{j=1}^{n} (y_{ij} - \bar{y}_{.j}) \right\}^2$$

$$= \frac{1}{kn^2} \left\{ \sum_{i=1}^{k} \sum_{j=1}^{n} (y_{ij} - \bar{y}_{.j})^2 + \sum_{i=1}^{k} \sum_{j \neq j'=1}^{n} (y_{ij} - \bar{y}_{.j})(y_{ij'} - \bar{y}_{.j'}) \right\} \tag{18}$$

The second term contains $kn(n-1)$ product terms, of which $2k(n-1)$ are products of y deviations from the respective strata means separated by one row, $2k(n-2)$ are products of y deviations separated by two rows, . . ., $2k(n-\alpha)$ are products of y deviations separated by α rows, . . ., and $2k$ separated by $(n-1)$ rows. We can, therefore, write (18) as

$$V(\bar{y}_{i.})_{sy} = \frac{1}{kn^2} \left\{ \sum_{i=1}^{k} \sum_{j=1}^{n} (y_{ij} - \bar{y}_{.j})^2 + 2 \sum_{i=1}^{k} \sum_{\alpha=1}^{n-1} \sum_{j=1}^{n-\alpha} (y_{ij} - \bar{y}_{.j})(y_{i\,j+\alpha} - \bar{y}_{.j+\alpha}) \right\}$$

$$= \frac{1}{kn^2} \left\{ \sum_{i=1}^{k} \sum_{j=1}^{n} (y_{ij} - \bar{y}_{.j})^2 + 2 \left(\sum_{\alpha=1}^{n-1} k(n-\alpha)\, \rho_{(\alpha)w} \right) \times \left(\sum_{j=1}^{n} \sum_{i=1}^{k} \frac{(y_{ij} - \bar{y}_{.j})^2}{n(k-1)} \right) \right\} \tag{19}$$

where

$$\rho_{(\alpha)w} = \frac{\displaystyle\sum_{i=1}^{k} \sum_{j=1}^{n-\alpha} \frac{(y_{ij} - \bar{y}_{.j})(y_{i\,j+\alpha} - \bar{y}_{.j+\alpha})}{k(n-\alpha)}}{\displaystyle\sum_{j=1}^{n} \sum_{i=1}^{k} \frac{(y_{ij} - \bar{y}_{.j})^2}{n(k-1)}} \tag{20}$$

and is termed the within stratum serial correlation between observations separated by distance ka. Hence, substituting from (17) in terms of S_{wr}^2, we obtain

$$V(\bar{y}_{i.})_{sy} = \frac{k-1}{k} \cdot \frac{S_{wr}^2}{n} \left\{1 + \frac{k}{k-1} \frac{2}{n} \sum_{a=1}^{n-1} (n-a)\, \rho_{(a)w}\right\} \quad (21)$$

The expression in this form was first derived by the Madows (1944). Comparing it with (16), we note that the relative efficiency of the systematic and stratified random samples depends on the values of $\rho_{(a)w}$ in the population, and no general conclusions can, therefore, be drawn. If the $\rho_{(a)w}$ are all positive, the stratified sample will be superior to the systematic sample; if the $\rho_{(a)w}$ are zero, the two samples will provide estimates of equal precision.

9.5 Comparison of Systematic with Simple and Stratified Random Samples for Certain Specified Populations

(i) *Linear Trend*

Let us suppose that the values of the successive units of the population increase in accordance with a linear law, so that

$$y_h = \mu + h \quad (22)$$

where μ is a constant and h goes from 1 to N.

Clearly,

$$\bar{y}_{..} = \frac{1}{N} \sum_{h=1}^{N} (\mu + h)$$

$$= \mu + \frac{N+1}{2} \quad (23)$$

$$\sum_{h=1}^{N} y_h^2 = N\mu^2 + \frac{N(N+1)(2N+1)}{6} + \mu N(N+1) \quad (24)$$

and hence

$$S^2 = \frac{1}{N-1}\left\{\sum_{h=1}^{N} y_h{}^2 - N\bar{y}_{..}{}^2\right\}$$

$$= \frac{nk\,(nk+1)}{12} \tag{25}$$

Similarly, since the observations within each row increase by unity, we have

$$S_{wr}{}^2 = \frac{k\,(k+1)}{12} \tag{26}$$

and for the same reason, since the column means corresponding to k different systematic samples also increase by unity, we have the mean square between column means given by

$$S_c{}^2 = \frac{k\,(k+1)}{12} \tag{27}$$

Substituting from (27) in (6), from (25) in (7), and from (26) in (16), we get

$$V_{S_y} = \frac{k^2-1}{12} \tag{28}$$

$$V_R = \frac{(k-1)\,(nk+1)}{12} \tag{29}$$

and

$$V_S = \frac{k^2-1}{12n} \tag{30}$$

Hence

$$V_S : V_{S_y} : V_R = \frac{k+1}{n} : k+1 : nk+1$$

or, approximately

$$\cong \frac{1}{n} : 1 : n \tag{31}$$

We notice that the variance of a stratified sample is only $1/n$-th the variance of a systematic sample, and the latter in its turn is also approximately $1/n$-th the variance of a random sample. Stratified sampling is thus seen to be the most efficient of the three methods for removing the effects of a linear trend, with systematic sampling following it as the next best method. The reader may like to verify that $\rho_{(a)w}$ is $(k-1)/k$ for all values of a, thus explaining the loss of efficiency of systematic sampling compared to stratified random sampling. He will also find that $\rho = -(k^2n+1)/(k^2n^2-1)$, or approximately $-1/n$, which accounts for the superiority of systematic sampling over random sampling.

(ii) *Periodic Variation*

We shall now consider populations in which sampling units with high and low values follow one another according to a regular pattern. Suppose such a population is represented by

$$y_h = \sin\left\{a + (h-1)\frac{\pi}{10}\right\}$$

where h varies from 1 to an integral multiple of 20. Clearly, the successive sampling units will repeat themselves after every 20th value. A systematic five per cent. sample from such a population will consist of sampling units drawn from the same position of each cycle, giving an estimate which is no more accurate than a single value. A five per cent. random sample, on the other hand, will contain units from different parts of the cycles with the result that the means of such samples will vary within a narrower range than the means of different systematic samples, thus making random samples more efficient than systematic samples for removing the effect of a periodic trend. At the other extreme, if we select a ten per cent. systematic sample with two regularly spaced observations from each cycle, the first selected randomly out of the first 10 and the second chosen at a distance of 10 units from the first, then the mean based thereon will be identical with the population mean, thus making systematic sampling the most efficient of all sampling methods. It will be noticed therefore that the relative efficiency of systematic sampling for populations showing periodic variation depends upon the choice of the interval between the successive units sought to be included in the sample. In particular, if the interval coincides with the period of the cycle, the sample will contain units which are all alike, giving $\rho = 1$, and, consequently, the relative increase in variance of systematic over random sampling is maximum. Of course, in nature, regular periodicity is most unlikely to occur but the example serves to illustrate how the effectiveness of a systematic sample is influenced by the interval in sampling a population exhibiting a periodic trend.

(iii) *Natural Populations*

Systematic sampling is found to be both efficient and convenient in sampling certain natural populations like forest areas for estimating the volume of timber (Hasel, 1942; Griffith, 1945–46) and areas under different types of cover (Osborne, 1942). We shall illustrate here its efficiency for sampling a certain natural population distributed in time.

Example 9.1

A pilot survey for investigating the possibility of estimating the catch of marine fish was conducted in a sample of landing centres on the Malabar coast of India (I.C.A.R., 1950). At each landing centre in the sample, a count was made of the number of boats landing every hour from 6 A.M. to 6 P.M. Out of the boats landing during each hour, the first one was selected for observation on weight of fish, the product with the number of landing boats giving an estimate of the catch brought during the hour. Table 9.1 shows the number of boats landing per hour at Quilandy Centre for seven consecutive Mondays. Calculate for each day the values of ρ between observations separated by $k = 2$, 3, 4 and 6 hours and hence investigate the effectiveness of systematic sampling relative to random sampling for making observations during the day.

The first step in the calculation consists in making the analysis of variance tables on the number of boaιs landing per hour for each of the several cases ($k = 2$, $n = 6$; $k = 3$, $n = 4$; $k = 4$, $n = 3$; and $k = 6$, $n = 2$) on the data for each day. The next step is to substitute from these tables the values of the mean squares between (nS_c^2) and within systematic samples (S_{wc}^2) in the expression for ρ, namely,

$$\rho = \frac{\frac{k-1}{k} S_c^2 - \frac{S_{wc}^2}{n}}{\frac{nk-1}{nk} S^2} \tag{32}$$

And finally we calculate the values of the variance of systematic relative to random sampling from (12). The calculations are illustrated in Table 9.2 with reference to data collected on the third Monday.

Table 9.3 presents the values of ρ and those of the variance of systematic relative to random sampling for all cases. It will be seen that ρ is negative for all except three cases, and smaller than $-1/11$, thus showing the superiority of systematic over random sampling for making observations. Further, it will be seen that the superiority generally improves with the size of the systematic sample.

TABLE 9.1

Number of Boats Landing during Each of 12 Hours (6 a.m. to 6 p.m.) on Seven Consecutive Mondays

Hour of Day / Week	6–7	7–8	8–9	9–10	10–11	11–12	12–1	1–2	2–3	3–4	4–5	5–6
1	42	52	19	6	23	56	36	59	14	14	2	6
2	23	39	33	27	52	32	39	45	33	9	6	0
3	31	25	13	6	15	9	19	2	9	31	16	27
4	41	14	15	24	19	24	21	27	21	32	45	57
5	8	16	14	8	47	39	20	26	33	34	61	45
6	2	6	20	28	28	17	40	41	41	37	21	31
7	16	15	4	14	12	4	15	10	17	10	18	16

TABLE 9.2

Analysis of Variance Table together with the Values of ρ and of the Variance of Systematic Relative to Random Sampling for the Data Collected on the Third Monday

Source of Variation	$k=6, n=2$		$k=4, n=3$		$k=3, n=4$		$k=2, n=6$	
	D.F.	Mean Square	D.F.	Mean Square	D.F.	Mean Square	D.F.	Mean Square
Between (nS_c^2) ..	5	47·08	3	52·97	2	70·08	1	0·75
Within (S_{wc}^2) ..	6	136·58	8	112·00	9	101·64	10	105·42
Total	11	95·90	11	95·90	11	95·90	11	95·90
ρ		−0·55		−0·27		−0·16		−0·199
$\% \frac{V_{Sy}}{V_R}$		49		55		73		1

TABLE 9.3

Values of the Intra-Class Coefficient of Correlation and of the Variance of Systematic Relative to Random Sampling

	Values of ρ				$\% \frac{V_{Sy}}{V_R}$			
Week \ Size of Sample n	2	3	4	6	2	3	4	6
1	+·34	−·25	−·26	−·13	147	61	30	63
2	−·33	−·36	−·18	−·16	74	33	63	37
3	−·55	−·27	−·16	−·199	49	55	73	1
4	−·29	−·29	−·31	−·19	78	51	8	12
5	+·51	−·48	+·04	−·193	166	5	154	7
6	−·79	−·27	−·32	−·197	23	55	7	3
7	−·58	−·23	−·16	−·131	46	67	73	63

9.6 Estimation of the Variance

Since a systematic sample is a random sample of one cluster only, no estimate of the variance can be formed from the sample. This is a great handicap of the method which otherwise offers great advantages.

Certain approximations are used in practice to calculate the variance. One of these consists in treating the systematic sample as if it were a random sample of n units and calculating the variance, using the formula

$$\left(\frac{1}{n} - \frac{1}{nk}\right) s_{wc}^2 \tag{33}$$

where s_{wc}^2 is the mean square between units within the selected systematic sample, *i.e.*, a column of diagram (1). Clearly, (33) does not provide an unbiased estimate of (6), for,

$$E\,(s_{wc}^2) = \frac{1}{n-1}\, E\left(\sum_{j=1}^{n} y_{ij}^2 - n\bar{y}_{i.}^2\right)$$

$$= \frac{1}{n-1}\left[\, E\left\{\sum_{j=1}^{n} y_{ij}^2\right\} - nE\,(\bar{y}_{i.}^2)\right]$$

$$= \frac{1}{n-1}\left[\frac{1}{k}\sum_{i=1}^{k}\sum_{j=1}^{n} y_{ij}^2 - n\,\{\bar{y}_{..}^2 + V\,(\bar{y}_{i.})_{sy}\}\right]$$

or, substituting from (11) in the right-hand side,

$$= \frac{1}{n-1}\left[\frac{1}{k}\sum_{i=1}^{k}\sum_{j=1}^{n} y_{ij}^2 - n\bar{y}_{..}^2 - n\,\frac{nk-1}{nk}\,\frac{\mathbf{S}^2}{n}\,\{1 + \rho\,(n-1)\}\right]$$

$$= \frac{nk-1}{nk}\,\mathbf{S}^2\,(1 - \rho) \tag{34}$$

Thus the expected value of (33) is given by

$$\left(\frac{1}{n} - \frac{1}{nk}\right) \frac{nk-1}{nk} \mathrm{S}^2 (1-\rho)$$

which clearly is different from the variance of the systematic sample given by (11), unless $\rho = -1/(nk-1)$, that is, unless the systematic sample behaves as a random sample of the population. It is only when the units in the population are randomly ordered that we may expect (33) to provide a satisfactory estimate of the variance of a systematic sample. If they are not randomly ordered, as will be the case in natural populations distributed in time and space, the effect of using (33) to estimate the variance will obviously be to under-estimate the variance on an average, if the intra-class correlation among units of the systematic sample is larger than $-1/(nk-1)$, and *vice versa*.

Another approximation sometimes used for calculating the variance from the sample is

$$\left(\frac{1}{n} - \frac{1}{nk}\right) \frac{\sum\limits_{j=1}^{n-1} (y_{ij} - y_{i\,j+1})^2}{2\,(n-1)} \tag{35}$$

This again will be a biased estimate, for we may write

$$\begin{aligned} E &\left\{\sum_{j=1}^{n-1} (y_{ij} - y_{i\,j+1})^2\right\} \\ &= E\left\{\sum_{j=1}^{n-1} (y_{ij} - \bar{y}_{.j} + \bar{y}_{.j} - y_{i\,j+1} + \bar{y}_{.j+1} - \bar{y}_{.j+1})^2\right\} \\ &= E\left[\sum_{j=1}^{n-1} \{(y_{ij} - \bar{y}_{.j}) - (y_{i\,j+1} - \bar{y}_{.j+1}) + (\bar{y}_{.j} - \bar{y}_{.j+1})\}^2\right] \\ &= E\left[\sum_{j=1}^{n-1} \{(y_{ij} - \bar{y}_{.j})^2 + (y_{i\,j+1} - \bar{y}_{.j+1})^2 + (\bar{y}_{.j} - \bar{y}_{.j+1})^2\right. \\ &\qquad - 2\,(y_{ij} - \bar{y}_{.j})\,(y_{i\,j+1} - \bar{y}_{.j+1}) \\ &\qquad + 2\,(\bar{y}_{.j} - \bar{y}_{.j+1})\,(y_{ij} - \bar{y}_{.j}) \\ &\qquad \left. - 2\,(\bar{y}_{.j} - \bar{y}_{.j+1})\,(y_{i\,j+1} - \bar{y}_{.j+1})\}\right] \end{aligned} \tag{36}$$

Clearly, the value of the first and the second term is equal to $(n-1)\,\mathrm{S}_{wr}^2$ each, that of the fourth is equal to $-2\,(n-1)\,\mathrm{S}_{wr}^2\rho_{(1)w}$, and that of the fifth and sixth is zero each. Hence we may write

$$\left(\frac{1}{n}-\frac{1}{nk}\right) E\left\{\frac{\sum\limits_{j=1}^{n-1}(y_{ij}-y_{i\,j+1})^2}{2\,(n-1)}\right\}$$

$$=\frac{k-1}{k}\cdot\frac{1}{n}\left\{\mathrm{S}_{wr}^2\,(1-\rho_{(1)w}) + \frac{1}{2\,(n-1)}\sum_{j=1}^{n-1}(\bar{y}_{.j}-\bar{y}_{.j+1})^2\right\} \qquad (37)$$

or, substituting from (21),

$$= V_{Sy} - 2\,\frac{\mathrm{S}_{wr}^2}{n^2}\sum_{a=1}^{n-1}(n-a)\,\rho_{(a)w} - \frac{k-1}{kn}\,\mathrm{S}_{wr}^2\rho_{(1)w} + \frac{k-1}{2kn\,(n-1)}\sum_{j=1}^{n-1}(\bar{y}_{.j}-\bar{y}_{.j+1})^2 \qquad (38)$$

It is difficult to put an easy interpretation on (38). We may say that if the differences between neighbouring rows are counterbalanced by the within stratum serial correlation, (35) may serve to give a fair idea of the variance of a systematic sample.

9.7 Two-Stage Sample: Equal Units: Systematic Sampling of Second-Stage Units

The method of systematic sampling can be used for the selection of a two-stage sample at either of the two stages or both. Of these schemes, of more interest is the one with systematic selection at the second stage, partly because this permits the estimation of the sampling error, and also because it enables the selection of the second-stage units to be entrusted to the field staff without great risk of errors and facilitates control of field work. The other two schemes do not possess the first advantage and for this reason will not be considered in the book, although their theory is straightforward.

We shall suppose that the population contains NM units, grouped into N first-stage units of M second-stage units each, and consider a scheme of sampling in which a sample of n first-stage units is selected with equal probabilities, and m second-stage units are selected from each one of the n first-stage units by the method of systematic sampling.

Following the previous notation for two-stage sampling, let $\bar{y}_{nm}$ denote the sample mean and $\bar{y}_{..}$ the population mean, defined by

$$\bar{y}_{nm} = \frac{1}{nm} \sum_{i}^{n} \sum_{j}^{m} y_{ij}$$

$$\bar{y}_{..} = \frac{1}{NM} \sum_{i=1}^{N} \sum_{j=1}^{M} y_{ij}$$

Clearly, $\bar{y}_{nm}$ provides an unbiased estimate of the population mean $\bar{y}_{..}$. For,

$$\begin{aligned} E(\bar{y}_{nm}) &= \frac{1}{n} E \left\{ \Sigma^{n} E(\bar{y}_{im} \mid i) \right\} \\ &= \frac{1}{n} E \left\{ \Sigma^{n} \bar{y}_{i.} \right\} \\ &= \bar{y}_{..} \end{aligned} \tag{39}$$

To obtain the variance of $\bar{y}_{nm}$, we write

$$\begin{aligned} V(\bar{y}_{nm}) &= E(\bar{y}_{nm} - \bar{y}_{..})^2 \\ &= E(\bar{y}_{nm} - \bar{y}_{n.} + \bar{y}_{n.} - \bar{y}_{..})^2 \\ &= E(\bar{y}_{nm} - \bar{y}_{n.})^2 + E(\bar{y}_{n.} - \bar{y}_{..})^2 \\ &\qquad + 2E(\bar{y}_{nm} - \bar{y}_{n.})(\bar{y}_{n.} - \bar{y}_{..}) \end{aligned} \tag{40}$$

Consider the first term in (40). We have

$$E(\bar{y}_{nm} - \bar{y}_{n.})^2 = \frac{1}{n^2} E \left[\Sigma^{n} (\bar{y}_{im} - \bar{y}_{i.}) \right]^2$$

$$= \frac{1}{n^2}\left[E\left\{\sum^{n}(\bar{y}_{im}-\bar{y}_{i.})^2 + \sum^{n}_{i\neq i'}(\bar{y}_{im}-\bar{y}_{i.})(\bar{y}_{i'm}-\bar{y}_{i'.})\right\}\right]$$

$$= \frac{1}{n^2}\left[E\sum^{n}E\{(\bar{y}_{im}-\bar{y}_{i.})^2 \mid i\} + E\sum^{n}_{i\neq i'}E\{(\bar{y}_{im}-\bar{y}_{i.})(\bar{y}_{i'm}-\bar{y}_{i'.}) \mid i, i'\}\right] \tag{41}$$

The value of the second term of (41) is clearly zero, since samples are independently drawn from the i-th and i'-th first-stage units. The value of the first term of (41) is derived from (11). We have

$$E\{(\bar{y}_{im}-\bar{y}_{i.})^2 \mid i\} = \left(1-\frac{1}{M}\right)\frac{S_i^2}{m}\{1+\rho_i(m-1)\} \tag{42}$$

where S_i^2 denotes the mean square between the second-stage units of the i-th selected first-stage unit, and ρ_i denotes the intra-class correlation between second-stage units within M/m columns of m units each which can be formed out of the M units of the i-th selected first-stage unit.

Substituting from (42) in (41), we then obtain

$$E(\bar{y}_{nm}-\bar{y}_{n.})^2 = \frac{1}{n^2}\left[E\sum\left(1-\frac{1}{M}\right)\frac{S_i^2}{m}\{1+\rho_i(m-1)\}\right]$$

$$= \frac{1}{nmN}\left(1-\frac{1}{M}\right)\sum_{i=1}^{N}S_i^2\{1+\rho_i(m-1)\} \tag{43}$$

The value of the second term in (40) is known to be given by

$$\left(\frac{1}{n}-\frac{1}{N}\right)S_b^2 \tag{44}$$

where

$$S_b^2 = \frac{1}{N-1}\sum_{i=1}^{N}(\bar{y}_{i.}-\bar{y}_{..})^2$$

while the last term is obviously zero. Interchanging the first and the second terms in (40), and substituting from (43) and (44), we thus obtain

$$V(\bar{y}_{nm}) = \left(\frac{1}{n} - \frac{1}{N}\right) S_b^2 + \frac{1}{nmN}\left(1 - \frac{1}{M}\right) \sum_{i=1}^{N} S_i^2 \{1 + \rho_i (m-1)\} \tag{45}$$

If S_i^2 is of the same order for all i, say equal to S_w^2, as is likely to be the case when the first-stage units are equal in size, we get

$$V(\bar{y}_{nm}) = \left(\frac{1}{n} - \frac{1}{N}\right) S_b^2 + \frac{1}{nm}\left(1 - \frac{1}{M}\right) S_w^2 \{1 + \bar{\rho}(m-1)\} \tag{46}$$

where

$$\bar{\rho} = \frac{1}{N} \sum_{i=1}^{N} \rho_i$$

Had the first-stage units been selected with replacement, equation (46) would be further simplified, giving

$$V(\bar{y}_{nm}) = \frac{S_b^2}{n} + \frac{1}{nm}\left(1 - \frac{1}{M}\right) S_w^2 \{1 + \bar{\rho}(m-1)\} \tag{47}$$

If $\rho_i = -1/(M-1)$, the method of systematic sampling for the selection of second-stage units will be equivalent to a method of random sampling, and we shall be left with the familiar expression for the variance of the mean of a two-stage random sample, namely,

$$V(\bar{y}_{nm}) = \left(\frac{1}{n} - \frac{1}{N}\right) S_b^2 + \frac{1}{n}\left(\frac{1}{m} - \frac{1}{M}\right) S_w^2 \tag{48}$$

To obtain the estimate of the variance, we consider the mean square between first-stage unit means in the sample, s_b^2, defined by

$$s_b^2 = \frac{1}{n-1} \sum^{n} (\bar{y}_{im} - \bar{y}_{nm})^2 \tag{49}$$

Taking expectations of both sides in (49), we may write

$$(n-1) \cdot E(s_b^2) = E\left(\sum_i^n \bar{y}_{im}^2\right) - nE(\bar{y}_{nm}^2) \tag{50}$$

Now to evaluate the first term in (50), we write

$$E\left(\sum^{n} \bar{y}_{im}{}^2\right) = E\left\{\sum^{n} E(\bar{y}_{im}{}^2 \mid i)\right\}$$

$$= E\left[\sum^{n} \{\bar{y}_{i.}{}^2 + V(\bar{y}_{im} \mid i)_{Sy}\}\right]$$

$$= E\left[\sum^{n} \left\{\bar{y}_{i.}{}^2 + \left(1 - \frac{1}{M}\right)\frac{\mathbf{S}_i{}^2}{m} \times \{1 + \rho_i(m-1)\}\right\}\right]$$

$$= \frac{n}{N}\sum_{i=1}^{N} \left\{\bar{y}_{i.}{}^2 + \left(1 - \frac{1}{M}\right)\frac{\mathbf{S}_i{}^2}{m} \times \{1 + \rho_i(m-1)\}\right\} \tag{51}$$

The value of the second term in (50) is directly obtained from (45), for,

$$nE(\bar{y}_{nm}{}^2) = n\{\bar{y}_{..}{}^2 + V(\bar{y}_{nm})\}$$

$$= n\left\{\bar{y}_{..}{}^2 + \left(\frac{1}{n} - \frac{1}{N}\right)\mathbf{S}_b{}^2 + \frac{1}{nm}\left(1 - \frac{1}{M}\right)\frac{1}{N} \times \sum_{i=1}^{N}\{1 + \rho_i(m-1)\}\,\mathbf{S}_i{}^2\right\} \tag{52}$$

Substituting from (51) and (52) in (50) and dividing by $(n - 1)$, we get

$$E(s_b{}^2) = \mathbf{S}_b{}^2 + \frac{1}{m}\left(1 - \frac{1}{M}\right)\frac{1}{N}\sum_{i=1}^{N}\{1 + \rho_i(m-1)\}\,\mathbf{S}_i{}^2 \tag{53}$$

We may, therefore, write from (45)

$$\text{Est. } V(\bar{y}_{nm}) = \frac{s_b{}^2}{n} \tag{54}$$

neglecting $\mathbf{S}_b{}^2/N$. In other words, $s_b{}^2/n$ supplies us with an upper bound for the estimate of the variance of the sample mean, which should be satisfactory for all practical purposes, particularly when N is large. When the first-stage units are selected with replacement, s_b^2/n gives an unbiased estimate of the sampling variance.

9.8 Two-Stage Sample: Unequal Units: Systematic Sampling of Second-Stage Units

In this section we shall extend the previous results to the case of unequal units. We shall suppose that the population consists of N first-stage units with the i-th unit containing M_i second-stage units ($i = 1, 2, \ldots, N$) and further consider a scheme of sampling in which the first-stage units are selected with replacement with varying probabilities $P_1, P_2, \ldots, P_N$ and the second-stage units within the selected first-stage units are selected by the method of systematic sampling. Let n denote the number of first-stage units to be selected in the sample and m_i the number of second-stage units to be selected from the i-th first-stage unit if included in the sample. Further, for simplicity, let m_i be so determined that M_i/m_i is an integer.

Following the treatment in Section 8.2, let

$$z_{ij} = \frac{M_i}{M_0}\frac{y_{ij}}{P_i} \tag{55}$$

and consider the estimate

$$\begin{aligned}\bar{z}_s &= \bar{z}_{n(m_i)} \\ &= \frac{1}{n}\sum^{n} \bar{z}_{i(m_i)}\end{aligned} \tag{56}$$

Clearly, $\bar{z}_s$ provides an unbiased estimate of the population mean $\bar{y}_{..}$. For,

$$\begin{aligned}E(\bar{z}_s) &= E\left\{\frac{1}{n}\sum^{n} \bar{z}_{i(m_i)}\right\} \\ &= E\left\{\frac{1}{n}\sum^{n} E(\bar{z}_{i(m_i)} \mid i)\right\} \\ &= E\left\{\frac{1}{n}\sum^{n} \bar{z}_{i.}\right\}\end{aligned}$$

in virtue of (4),

$$= \sum_{i=1}^{N} P_i \bar{z}_{i.}$$

$$= \bar{y}_{..} \tag{57}$$

To obtain the sampling variance of $\bar{z}_S$, we write

$$\begin{aligned} V(\bar{z}_s) &= E(\bar{z}_s - \bar{z}_{..})^2 \\ &= E(\bar{z}_{n(m_i)} - \bar{z}_{n.} + \bar{z}_{n.} - \bar{z}_{..})^2 \\ &= E(\bar{z}_{n(m_i)} - \bar{z}_{n.})^2 + E(\bar{z}_{n.} - \bar{z}_{..})^2 \\ &\qquad + 2E\{(\bar{z}_{n(m_i)} - \bar{z}_{n.})(\bar{z}_{n.} - \bar{z}_{..})\} \end{aligned} \tag{58}$$

Taking the first term in (58), we have

$$\begin{aligned} E(\bar{z}_{n(m_i)} - \bar{z}_{n.})^2 &= \frac{1}{n^2} E\left\{\sum^{n} (\bar{z}_{i(m_i)} - \bar{z}_{i.})\right\}^2 \\ &= \frac{1}{n^2} E\left\{\sum^{n} (\bar{z}_{i(m_i)} - \bar{z}_{i.})^2 \right. \\ &\qquad \left. + \sum_{i \neq i'}^{n} (\bar{z}_{i(m_i)} - \bar{z}_{i.})(\bar{z}_{i'(m_{i'})} - \bar{z}_{i'.})\right\} \\ &= \frac{1}{n^2} E\left[\sum^{n} E\{(\bar{z}_{i(m_i)} - \bar{z}_{i.})^2 \mid i\}\right. \\ &\qquad + \sum_{i \neq i'}^{n} E\{(\bar{z}_{i(m_i)} - \bar{z}_{i.}) \mid i\} \\ &\qquad \left. \times E\{(\bar{z}_{i'(m_{i'})} - \bar{z}_{i'.}) \mid i'\}\right] \end{aligned}$$

The expression under the summation sign in the first term represents the variance of the mean of a systematic sample of m_i selected out of M_i and can be written from (42). The value of the second term is clearly zero. We, therefore, write

$$\begin{aligned} E(\bar{z}_{n(m_i)} - \bar{z}_{n.})^2 &= \frac{1}{n^2} E\left[\sum^{n} V(\bar{z}_{i(m_i)} \mid i)_{Sy}\right] \\ &= \frac{1}{n^2} E\left[\sum^{n} \left(1 - \frac{1}{M_i}\right) \frac{\mathbf{S}_i^2}{m_i} \{1 + \rho_i(m_i - 1)\}\right] \\ &= \frac{1}{n} \sum_{i=1}^{N} P_i \left(1 - \frac{1}{M_i}\right) \frac{\mathbf{S}_i^2}{m_i} \{1 + \rho_i(m_i - 1)\} \end{aligned}$$

The value of the second term in (58) is known to be σ_{bz}^2/n, where

$$\sigma_{bz}^2 = \sum_{i=1}^{N} P_i\,(\bar{z}_{i.} - \bar{z}_{..})^2$$

The last term in (58) is obviously zero. Hence we have

$$V(\bar{z}_s) = \frac{\sigma_{bz}^2}{n} + \frac{1}{n}\sum_{i=1}^{N} P_i\left(1 - \frac{1}{M_i}\right)\frac{S_i^2}{m_i}\,\{1 + \rho_i\,(m_i - 1)\} \tag{59}$$

To obtain an estimate of $V(\bar{z}_s)$, we consider the mean square between the first-stage unit means in the sample, namely,

$$s_{bz}^2 = \frac{1}{n-1}\sum^{n} (\bar{z}_{i(m_i)} - \bar{z}_{n(m_i)})^2 \tag{60}$$

Expanding and taking expectations, we write from (17) of Section 8.3,

$$E(s_{bz}^2) = \frac{1}{n-1}\Big[n\,\{V(\bar{z}_{i(m_i)})_{sy} + \bar{z}_{..}^2\} - n\,\{V(\bar{z}_{n(m_i)}) + \bar{z}_{..}^2\}\Big] \tag{61}$$

where $V(\bar{z}_{i(m_i)})_{sy}$ is the variance of the estimate based on a sample of one first-stage unit selected with probability P_i and m_i second-stage units selected by the method of systematic sampling therefrom. Substituting from (59) in (61), we therefore get

$$E(s_{bz}^2) = nV(\bar{z}_s)$$

whence

$$\text{Est. } V(\bar{z}_s) = \frac{s_{bz}^2}{n} \tag{62}$$

showing that s_{bz}^2/n supplies us with an unbiased estimate of the variance of $\bar{z}_s$ for the scheme of two-stage sampling considered in this section.

Example 9.2

Reference has been made in Example 9.1 to a pilot survey for estimating the total catch of marine fish conducted along a

100-mile strip of Malabar coast during 1950. The landing centres along the coast were first divided into two groups: (*a*) those known for their high fishing activity, and (*b*) all the rest. From the latter stratum, comprising 59 centres, a sample of 10 centres was selected with replacement with probability proportional to the number of fishing boats as enumerated at the previous census. The number of boats for all the 59 centres was 4573. Table 9.4 gives for the selected centres the number of boats at the previous census and the estimated average catch per hour for a certain day of the survey based on each of two systematic samples of $m = 6$ and $m = 2$ hours. Make an estimate of the average catch per hour for the stratum and calculate its standard error.

It is proposed to extend the survey to the entire Indian coast. Use the Malabar experience to determine the number of centres required for estimating the daily catch with, say, 5% standard error.

Let A_i be the number of boats at the i-th centre and $A = \sum_{i=1}^{N} A_i = 4573$ the total number of boats for the stratum. Clearly then the selection probability P_i for the i-th centre is given by $P_i = A_i/A$. Further, let $\bar{y}_{im}$ be the average catch per hour at the i-th centre based on a sample of m hours. Then the estimated average catch per hour for the entire stratum and its standard error are respectively given by $\bar{z}_{nm}$ and $s_{bz}/\sqrt{n}$, where

$$\bar{z}_{nm} = \frac{1}{n} \sum^{n} \bar{z}_{im}$$

and

$$s_{bz}{}^2 = \frac{1}{n-1} \sum^{n} (\bar{z}_{im} - \bar{z}_{nm})^2$$

with

$$\bar{z}_{im} = \frac{\bar{y}_{im}}{NP_i} = \frac{A}{59A_i}\, \bar{y}_{im}$$

The values of $\bar{z}_{im}$ for the selected centres are given in the last two columns of Table 9.4, and the various steps in the computation of the average catch per hour and its standard error are shown below the table. It is seen that for $m = 6$, the estimated average catch per hour works out to 735·8 mds. with a standard error of 164·9; and for $m = 2$, the average catch per hour works out to 759·1 mds. with a standard error of 220·8. Expressed as percentage of the estimated mean, the standard errors are 22% and 29% respectively.

TABLE 9.4

The Number of Enumerated Boats and the Estimated Average Catch in Mds./Hr. for a Sample of 10 Landing Centres along the Malabar Coast

Serial No. of the Centre	No. of Boats (A_i)	Estimated Average Catch (Mds./Hr.) ($\bar{y}_{im}$)		$\bar{z}_{im} = \frac{A}{59A_i}\bar{y}_{im}$	
		$m = 6$	$m = 2$	$m = 6$	$m = 2$
(1)	(2)	(3)	(4)	(5)	(6)
1	68	387·50	729·50	441·68	831·51
2	36	840·17	1123·00	1808·90	2417·83
3	96	1094·17	754·00	883·41	608·76
4	45	61·67	55·00	106·22	94·73
5	103	899·00	1072·00	676·51	806·69
6	74	1401·67	1277·00	1468·13	1337·54
7	127	677·33	162·50	413·38	99·17
8	174	1331·17	634·00	592·97	282·42
9	12	76·50	54·50	494·12	352·02
10	18	109·83	176·50	472·93	760·01
(1) Total $= \Sigma \bar{z}_{im}$				7358·25	7590·68
(2) Estimated average catch (Mds./Hr.) $= \bar{z}_{nm}$				735·83	759·07
(3) Correction factor $= \frac{1}{10}(\Sigma \bar{z}_{im})^2$ ..				5414384·31	5761842·29
(4) Crude S.S. $= \Sigma(\bar{z}_{im}^2)$				7862281·63	10147764·33
(5) Adjusted S.S. $= \Sigma(\bar{z}_{im}^2) - \frac{1}{10}(\Sigma \bar{z}_{im})^2$				2447897·32	4385922·04
(6) Mean square $= s_{bz}^2$				271988·59	487324·67
(7) Variance $(\bar{z}_{nm}) = \frac{1}{10} s_{bz}^2$				27198·86	48732·47
(8) Standard error (Mds.)				164·9	220·8
(9) Standard error as % of estimated average				22	29
(10) No. of centres for 5% *S.E.*				201	339

The number of centres required for estimating the daily catch along the entire coast with 5% standard error is given by

$$\hat{n} = \frac{s_{b\bar{z}}^{2}}{\cdot 0025\bar{\bar{z}}_{nm}^{2}}$$

$$= 201 \qquad \text{for } m = 6$$

$$= 339 \qquad \text{for } m = 2$$

REFERENCES

1. Madow, W. G. and L. H. (1944) "On the Theory of Systematic Sampling," *Ann. Math. Statist.*, **15**, 1–24.

2. Madow, L. H. (1946) .. "Systematic Sampling and its Relation to Other Sampling Designs," *Jour. Amer. Statist. Assoc.*, **41**, 204–17.

3. Hasel, A. A. (1942) .. "Estimation of Volume in Timber Stands by Strip Sampling," *Ann. Math. Statist.*, **13**, 179–206.

4. Osborne, J. G. (1942) .. "Sampling Errors of Systematic and Random Surveys of Cover-Type Areas," *Jour. Amer. Statist. Assoc.*, **37**, 256–64.

5. I.C.A.R., New Delhi (1950) "Report on the Pilot Sample Survey Conducted on the Malabar Coast for Estimating the Catch of Marine Fish." *Unpublished.*

6. Griffith, A. L. (1945–46) *The Efficiency of Enumerations.* Forest Research Institute, Dehra Dun, Indian Forest Leaflets, Nos. 83–93.

CHAPTER X

NON-SAMPLING ERRORS

A. OBSERVATIONAL ERRORS

10a.1 Introduction

In developing the sampling theory in the preceding chapters, we assumed that the character observed on the i-th unit of the population ($i = 1, 2, \ldots, N$), takes a unique value y_i whenever the unit is included in the sample, irrespective of the person who enumerates it. By implication we assumed that a complete count of all the N units gives a unique value for the mean or the total of the population. In practice, however, the situation is rarely so simple as the one described above, since the value observed on any unit will also depend upon the enumerator reporting the value. Thus, an eye-estimate of the yield of a crop in a field will depend upon the judgment of the enumerator making the estimate and will invariably be different from the true value of the yield obtained by harvesting the crop in the field. The magnitude and direction of the difference will depend upon the enumerator's intrinsic tendencies or biases and the approach to the selected unit or interviewee at the time of reporting the value. Even with factual characters like those of farm facts, *e.g.*, the area under crops, the number of animals on the farm, etc., there is found to be a marked variation in the performance of the same or different enumerators. It follows that even when the sampling fraction is unity, or, in other words, a complete count of all the N units is made, the result will vary in repeated counts. As the errors responsible for this variation arise in the process of collecting data, they are properly *observational errors* but are also referred to as *response errors* (Hansen *et al.*, 1951). Together with the errors arising from incomplete samples and faulty procedures of estimation, they go to make up what are termed as *non-sampling errors.*

We have given several examples in Chapter I to show that the net effect of non-sampling errors on the value of the estimate can sometimes be serious, and that it is, therefore, important to control them as far as practicable. In Part A of this chapter we shall deal with the measurement and control of observational errors and in Part B with the treatment of incomplete samples.

10*a*.2 Mathematical Model for the Measurement of Observational Errors

Let x_i $(i = 1, 2, \ldots, h)$, denote the true value of the character on the i-th unit in a simple random sample of h units drawn from N units, and y_{ijk},

$$i = 1, 2, \ldots, h$$

$$j = 1, 2, \ldots, m$$

$$k = 0, 1, 2, \ldots, n_{ij}$$

denote the value reported by the j-th enumerator on the i-th unit for the k-th occasion. It will be seen that m enumerators have been assumed to participate in the survey, with the j-th enumerator making n_{ij} observations on the i-th unit in the sample.

The difference between the reported value and the true value is called the *error of observation*, and for any given measurement technique will depend upon the enumerator reporting the value, the interaction of the enumerator with the true value of the unit, and the mood and like causes at the time of reporting. The reported value may, therefore, be considered as being made up of four uncorrelated components as follows:

$$y_{ijk} = x_i + a_j + \delta_{ij} + \epsilon'_{ijk} \tag{1}$$

where

a_j represents the bias of the j-th enumerator in repeated observations on all units,

δ_{ij} the interaction of the j-th enumerator with the i-th unit,

ϵ'_{ijk} the deviation from $x_i + a_j + \delta_{ij}$ when the j-th enumerator reports on the i-th unit on the k-th occasion, Equations (4) and (5) are now simplified, being given by

$$\bar{y}_{.j} = \frac{1}{\bar{\bar{n}}} \sum_{i}^{h} x_i n_{ij} + a_j + \frac{1}{\bar{\bar{n}}} \sum_{i}^{h} \epsilon_{ij} n_{ij} \tag{6}$$

and

$$\bar{y}_{..} = \frac{1}{h} \sum_{i}^{h} x_i + \frac{1}{m} \sum_{j}^{m} a_j + \frac{1}{n} \sum_{j}^{m} \sum_{i}^{h} \epsilon_{ij} n_{ij} \tag{7}$$

It follows that

$$\begin{aligned} E(\bar{y}_{.j}) &= \frac{1}{N} \sum_{i=1}^{N} x_i + \frac{1}{M} \sum_{j=1}^{M} a_j \\ &= \mu + \bar{a} \end{aligned} \tag{8}$$

where μ is the population mean of the true values, to be estimated, and $\bar{a}$ is the population mean of enumerators' biases. Also,

$$E(\bar{y}_{..}) = \mu + \bar{a} \tag{9}$$

It will be seen that the sample mean $\bar{y}_{..}$ does not provide an unbiased estimate of μ, unless a_j's vary in such a way that $\bar{a}$ is zero. Experience indicates that although a_j is usually variable from enumerator to enumerator, $\bar{a}$ is not always negligible. Thus in estimating the crop by eye, there is a tendency to under-estimate the crop in good years and over-estimate it in bad years. Consequently, in a good year the bulk of the enumerators under-estimate the crop, resulting in a significant negative value for $\bar{a}$, and in a bad year the bulk of them over-estimate it giving $\bar{a}$ a significant positive value. Again, as pointed out in Chapter I, when a crop is unevenly sown as in India, sample-harvesting it by small plots, such as those marked with a portable frame, may result in over-estimation of yield (Sukhatme, 1947). It is, therefore, of the highest importance in surveys to ensure that the bias $\bar{a}$ is negligible.

(c) *Variance of the Sample Mean*

By definition,

$$V(\bar{y}_{.j}) = E\{\bar{y}_{.j} - E(\bar{y}_{.j})\}^2$$

Substituting for $\bar{y}_{.j}$ and $E(\bar{y}_{.j})$ from (6) and (8), we have

$$V(\bar{y}_{.j}) = E\left\{\frac{1}{\bar{n}}\sum_{i}^{h} x_i n_{ij} - \mu + a_j - \bar{a} + \frac{1}{\bar{n}}\sum_{i}^{h} \epsilon_{ij} n_{ij}\right\}^2 \qquad (10)$$

On squaring the expression within brackets in (10), taking expectations term by term, and noting that the expectations of the product terms are zero, we have

$$V(\bar{y}_{.j}) = E\left(\frac{1}{\bar{n}}\sum_{i}^{h} x_i n_{ij} - \mu\right)^2 + E(a_j - \bar{a})^2 + \frac{1}{\bar{n}^2} E\left(\sum_{i}^{h} \epsilon_{ij} n_{ij}\right)^2 \qquad (11)$$

To evaluate the first term in (11), we write

$$\begin{aligned} E\left(\frac{1}{\bar{n}}\sum_{i}^{h} x_i n_{ij} - \mu\right)^2 &= \frac{1}{\bar{n}^2} E\left(\sum_{i}^{h} x_i^2 n_{ij}^2 + \sum_{i \neq i'}^{h} x_i x_{i'} n_{ij} n_{i'j}\right) - \mu^2 \\ &= \frac{1}{\bar{n}^2}\left\{\sum_{i}^{h} E(x_i^2)\, n_{ij}^2 + \sum_{i \neq i'}^{h} E(x_i x_{i'})\, n_{ij} n_{i'j}\right\} - \mu^2 \end{aligned}$$

Substituting for $E(x_i^2)$ and $E(x_i x_{i'})$ from (21) and (34) of Chapter II, we have

$$\begin{aligned} E\left(\frac{1}{\bar{n}}\sum_{i}^{h} x_i n_{ij} - \mu\right)^2 &= \frac{1}{\bar{n}^2}\left\{\sum_{i}^{h}\left(\frac{N-1}{N}\mathbf{S}_x^2 + \mu^2\right) n_{ij}^2 + \sum_{i \neq i'}^{h}\left(\mu^2 - \frac{\mathbf{S}_x^2}{N}\right) n_{ij} n_{i'j}\right\} - \mu^2 \\ &= \mathbf{S}_x^2\left(\frac{1}{\bar{n}^2}\sum_{i}^{h} n_{ij}^2 - \frac{1}{N}\right) \\ &= \mathbf{S}_x^2\left(\frac{1}{\bar{n}} - \frac{1}{N}\right) \end{aligned} \qquad (12)$$

where

$$\mathbf{S}_x^2 = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \mu)^2 \tag{13}$$

The value of the second term in (11) is given by

$$E\,(\alpha_j - \bar{\alpha})^2 = \mathbf{S}_\alpha^2 \left(1 - \frac{1}{M}\right) \tag{14}$$

where

$$\mathbf{S}_\alpha^2 = \frac{1}{M-1} \sum_{j=1}^{M} (\alpha_j - \bar{\alpha})^2 \tag{15}$$

while that of the third term in (11), by

$$\begin{aligned} \frac{1}{\bar{n}^2}\, E\left\{\left(\sum_{i}^{h} \epsilon_{ij} n_{ij}\right)^2 \Bigg|\, j\right\} &= \frac{1}{\bar{n}^2}\, E\, \left\{\sum_{i}^{h} \epsilon_{ij}^2 n_{ij}^2 + \sum_{i \neq i'}^{h} \epsilon_{ij}\epsilon_{i'j} n_{ij} n_{i'j}\right\} \\ &= \frac{1}{\bar{n}^2}\, \left\{\sum_{i}^{h} E\,(\epsilon_{ij}^2 \mid j)\, n_{ij}^2 \right. \\ &\qquad \left. + \sum_{i \neq i'}^{h} E\,(\epsilon_{ij}\epsilon_{i'j} \mid j)\, n_{ij} n_{i'j}\right\} \\ &= \frac{1}{\bar{n}^2}\, \left(\mathbf{S}_\epsilon^2 \sum_{i}^{h} n_{ij}^2\right) \\ &= \frac{\mathbf{S}_\epsilon^2}{\bar{n}} \end{aligned} \tag{16}$$

since

$$E\,(\epsilon_{ij}^2 \mid j) = \mathbf{S}_{\epsilon_j}^2 = \mathbf{S}_\epsilon^2$$

and

$$E\,(\epsilon_{ij}\epsilon_{i'j}) = 0$$

Hence, substituting from (12), (14) and (16) in (11), we get

$$V\,(\bar{y}_{.j}) = \mathbf{S}_x^2 \left(\frac{1}{\bar{n}} - \frac{1}{N}\right) + \mathbf{S}_\alpha^2 \left(1 - \frac{1}{M}\right) + \frac{1}{\bar{n}}\, \mathbf{S}_\epsilon^2 \tag{17}$$

For N and M infinitely large, we have

$$V\,(\bar{y}_{.j}) = \frac{1}{\bar{n}}\, (\mathbf{S}_x^2 + \mathbf{S}_\epsilon^2) + \mathbf{S}_\alpha^2 \tag{18}$$

To obtain the variance of $\bar{y}_{..}$, we proceed in a similar way. We write

$$\begin{aligned}
V(\bar{y}_{..}) &= E\{\bar{y}_{..} - E(\bar{y}_{..})\}^2 \\
&= E\left\{\frac{1}{h}\sum_{i}^{h} x_i - \mu + \frac{1}{m}\sum_{j}^{m} a_j - \bar{a}\right. \\
&\qquad\qquad \left. + \frac{1}{n}\sum_{j}^{m}\sum_{i}^{h} \epsilon_{ij} n_{ij}\right\}^2 \\
&= E\left(\frac{1}{h}\sum_{i}^{h} x_i - \mu\right)^2 + E\left(\frac{1}{m}\sum_{j}^{m} a_j - \bar{a}\right)^2 \\
&\qquad\qquad + \frac{1}{n^2} E\left(\sum_{j}^{m}\sum_{i}^{h} \epsilon_{ij} n_{ij}\right)^2 \qquad (19)
\end{aligned}$$

Now, since $x_1, x_2, \ldots, x_h$ is a simple random sample of the population,

$$E\left(\frac{1}{h}\sum_{i}^{h} x_i - \mu\right)^2 = \mathrm{S}_x^2\left(\frac{1}{h} - \frac{1}{N}\right) \qquad (20)$$

Similarly,

$$E\left(\frac{1}{m}\sum_{j}^{m} a_j - \bar{a}\right)^2 = \mathrm{S}_a^2\left(\frac{1}{m} - \frac{1}{M}\right) \qquad (21)$$

It remains to evaluate the third term only. We have

$$\begin{aligned}
\frac{1}{n^2} E\left(\sum_{j}^{m}\sum_{i}^{h} \epsilon_{ij} n_{ij}\right)^2 &= \frac{1}{n^2} E\left\{\sum_{j}^{m}\left(\sum_{i}^{h} \epsilon_{ij} n_{ij}\right)^2\right\} \\
&\quad + \frac{1}{n^2} E\left\{\sum_{j \neq j'}^{m}\left(\sum_{i}^{h} \epsilon_{ij} n_{ij}\right)\right. \\
&\qquad \left. \times\left(\sum_{i}^{h} \epsilon_{ij'} n_{ij'}\right)\right\}
\end{aligned}$$

$$= \frac{1}{n^2} E\left[\sum_j^m E\left\{\left(\sum_i^h \epsilon_{ij} n_{ij}\right)^2 \mid j\right\}\right]$$

$$+ \frac{1}{n^2} E\left\{\sum_{j \neq j'}^m E\left(\sum_i^h \epsilon_{ij} n_{ij} \mid j\right)\right.$$

$$\left. \times E\left(\sum_i^h \epsilon_{ij'} n_{ij'} \mid j'\right)\right\}$$

$$= \frac{\mathbf{S}_\epsilon^2}{n}$$

$$= \frac{\mathbf{S}_\epsilon^2}{hp} \tag{22}$$

since

$$E\left(\sum_i^h \epsilon_{ij} n_{ij} \mid j\right) = 0 = E\left(\sum_i^h \epsilon_{ij'} n_{ij'} \mid j'\right)$$

Hence, substituting from (20), (21) and (22) in (19), we obtain

$$V(\bar{y}_{..}) = \mathbf{S}_x^2\left(\frac{1}{h} - \frac{1}{N}\right) + \mathbf{S}_\alpha^2\left(\frac{1}{m} - \frac{1}{M}\right) + \frac{\mathbf{S}_\epsilon^2}{hp} \tag{23}$$

For N and M infinitely large, we have

$$V(\bar{y}_{..}) = \frac{\mathbf{S}_x^2}{h} + \frac{\mathbf{S}_\alpha^2}{m} + \frac{\mathbf{S}_\epsilon^2}{hp} \tag{24}$$

Further, if $p = 1$, which will usually be the case in practice, the expression for $V(\bar{y}_{..})$ will be simplified, being given by

$$V(\bar{y}_{..}) = \frac{\mathbf{S}_x^2 + \mathbf{S}_\epsilon^2}{h} + \frac{\mathbf{S}_\alpha^2}{m} \tag{25}$$

which can be alternatively expressed as

$$= \frac{\mathbf{S}_x^2 + \mathbf{S}_\alpha^2 + \mathbf{S}_\epsilon^2}{h} + \mathbf{S}_\alpha^2\left(\frac{1}{m} - \frac{1}{h}\right)$$

$$= \frac{\mathbf{S}_y^2}{h} + \mathbf{S}_\alpha^2\left(\frac{1}{m} - \frac{1}{h}\right) \tag{26}$$

since the variance of a single observation drawn from an infinite population when M is infinitely large is clearly $\mathbf{S}_x^2 + \mathbf{S}_\alpha^2 + \mathbf{S}_\epsilon^2 = \mathbf{S}_y^2$.

The above formulæ must be considered as fundamental formulæ in the theory of sample surveys. They show that the sampling variance of the estimate is not entirely due to errors arising from chance variation in selection of the sample of h units, but is inflated by the variability in biases of the enumerators. Consequently, the formulæ given in the previous chapters under-estimate the actual sampling variance to which the estimates are subject. This emphasises another point that it is not sufficient in surveys to ensure that the α_j's cancel each other on the average, but that it is necessary also to see that the effect of variability in the α_j's is reduced to the minimum, by taking the maximum care in planning details of the surveys such as the questionnaire, the method of making observations, the training of field staff and the supervision over their work.

It follows too that the variance of the sample estimate does not reduce to zero even when a complete count is made. For, the limit of equation (26) when $h = N$ and N becomes infinitely large is given by

$$V(\bar{y}_{..} \mid h = \infty) = \frac{S_\alpha'^2}{m'}$$

where $m' = N/\bar{n}$ is the number of enumerators required for the complete count and selected out of a population of M' enumerators assumed infinitely large, and where $S_\alpha'^2$ represents the variability of the biases of the population of enumerators under census conditions. This expression provides us with an important consideration in determining the relative roles of complete enumeration and random sampling. In particular, we notice that a sample may give a more reliable estimate than a census when

$$\frac{S_\alpha'^2}{m'} > \frac{S_y^2}{h} + S_\alpha^2\left(\frac{1}{m} - \frac{1}{h}\right)$$

It is in fact in this possibility of controlling the magnitude of the variability among the α_j's in a relatively small survey compared to a census by adopting more refined methods of enumeration, not always possible in a census, and by recruiting better-trained and better-paid enumerators, that lies the effective role of sampling for collecting information.

10α.4 Estimation of the Different Components of Variance

Let s_e^2 denote the mean square between the means of m enumerators, defined by

$$s_e^2 = \frac{1}{m-1} \sum_{j}^{m} (\bar{y}_{.j} - \bar{y}_{..})^2 \tag{27}$$

Then expanding and taking expectations, we write

$$\begin{aligned} (m-1)\cdot E\,(s_e^2) &= \sum_{j}^{m} E\,(\bar{y}_{.j}^2) - mE\,(\bar{y}_{..}^2) \\ &= \sum_{j}^{m} [V(\bar{y}_{.j}) + \{E(\bar{y}_{.j})\}^2] - m\,[V(\bar{y}_{..}) + \{E(\bar{y}_{..})\}^2] \\ &= \sum_{j}^{m} \{V\,(\bar{y}_{.j}) + (\mu + \bar{a})^2\} - m\,\{V(\bar{y}_{..}) + (\mu + \bar{a})^2\} \\ &= \sum_{j}^{m} V\,(\bar{y}_{.j}) - m\,V(\bar{y}_{..}) \end{aligned} \tag{28}$$

Substituting from (17) and (23) in (28), we get

$$E\,(s_e^2) = \mathbf{S}_x^2 \cdot \frac{m}{hp} \cdot \frac{m-p}{m-1} + \mathbf{S}_\alpha^2 + \frac{m}{hp}\,\mathbf{S}_\epsilon^2 \tag{29}$$

Similarly, denoting by s_{eo}^2 the mean square between observations within enumerators, defined by

$$s_{eo}^2 = \frac{1}{n-m} \sum_{j}^{m} \sum_{i}^{\bar{n}} (y_{ij} - \bar{y}_{.j})^2 \tag{30}$$

we get

$$\begin{aligned} (n-m)\cdot E\,(s_{eo}^2) &= E\left\{\sum_{j}^{m} \sum_{i}^{\bar{n}} y_{ij}^2 - \bar{n} \sum_{j}^{m} \bar{y}_{.j}^2\right\} \\ &= \sum_{j}^{m} \sum_{i}^{\bar{n}} V\,(y_{ij}) - \bar{n} \sum_{j}^{m} V\,(\bar{y}_{.j}) \end{aligned} \tag{31}$$

Now putting $\bar{n} = 1$ in (17), we have

$$V\,(y_{ij}) = \mathbf{S}_x^2\left(1 - \frac{1}{N}\right) + \mathbf{S}_\alpha^2\left(1 - \frac{1}{M}\right) + \mathbf{S}_\epsilon^2 \tag{32}$$

Hence, substituting from (17) and (32) in (31), we get

$$E\,(s_{eo}^2) = \mathbf{S}_x^2 + \mathbf{S}_\epsilon^2 \tag{33}$$

Equations (29) and (33) must be supplemented by additional information in order to be able to estimate $\mathbf{S}_x^2$, $\mathbf{S}_\epsilon^2$ and $\mathbf{S}_\alpha^2$ separately. We shall, therefore, consider yet another mean square, namely, that between unit means, s_u^2, given by

$$s_u^2 = \frac{1}{h-1} \sum_i^h (\bar{y}_{i.} - \bar{y}_{..})^2 \tag{34}$$

Expanding and taking expectations, we write

$$\begin{aligned}(h-1) \cdot E\,(s_u^2) &= \sum_i^h E\,(\bar{y}_{i.}^2) - hE\,(\bar{y}_{..}^2)\\ &= \sum_i^h \{V\,(\bar{y}_{i.})\} - h\,V\,(\bar{y}_{..})\end{aligned} \tag{35}$$

The number of enumerators making observations on any unit is p. Hence, from (23), we get

$$V\,(\bar{y}_{i.}) = \mathbf{S}_x^2 \left(1 - \frac{1}{N}\right) + \mathbf{S}_\alpha^2 \left(\frac{1}{p} - \frac{1}{M}\right) + \frac{1}{p}\,\mathbf{S}_\epsilon^2 \tag{36}$$

Substituting from (36) and (23) in (35), we obtain

$$E\,(s_u^2) = \mathbf{S}_x^2 + \frac{1}{p}\,\mathbf{S}_\epsilon^2 + \frac{h}{h-1} \cdot \frac{m-p}{mp}\,\mathbf{S}_\alpha^2 \tag{37}$$

The set of three equations (29), (33) and (37) provides the estimates of $\mathbf{S}_x^2$, $\mathbf{S}_\alpha^2$ and $\mathbf{S}_\epsilon^2$. In particular, we obtain

$$\text{Est. } \mathbf{S}_\alpha^2 = \frac{p\,(m-1)\,(h-1)}{pmh - ph - pm + m} \left\{ s_e^2 + \frac{m}{h\,(m-1)}\,s_u^2 - \frac{m^2}{hp\,(m-1)}\,s_{eo}^2 \right\} \tag{38}$$

In practice, however, as already noted, p will be 1. The equations (29), (33) and (37) then simplify as follows:

$$E(s_e^2) = \frac{m}{h}(\mathbf{S}_x^2 + \mathbf{S}_\epsilon^2) + \mathbf{S}_\alpha^2 \tag{39}$$

$$E(s_{eo}^2) = \mathbf{S}_x^2 + \mathbf{S}_\epsilon^2 \tag{40}$$

and

$$E(s_u^2) = \mathbf{S}_x^2 + \mathbf{S}_\epsilon^2 + \frac{h}{h-1} \cdot \frac{m-1}{m} \mathbf{S}_\alpha^2 \tag{41}$$

The three equations are no longer linearly independent. For, when $p = 1$, the three mean squares are connected by the identity:

$$(h-1)\, s_u^2 \equiv (m-1)\frac{h}{m}\, s_e^2 + (h-m)\, s_{eo}^2 \tag{42}$$

We further notice that $\mathbf{S}_x^2 + \mathbf{S}_\epsilon^2$ occurs together in all the equations. Solving then for $\mathbf{S}_\alpha^2$ and $\mathbf{S}_x^2 + \mathbf{S}_\epsilon^2$ together any two of the three equations, we obtain

$$\text{Est. } \mathbf{S}_\alpha^2 = s_e^2 - \frac{m}{h}\, s_{eo}^2 \tag{43}$$

and

$$\text{Est. } (\mathbf{S}_x^2 + \mathbf{S}_\epsilon^2) = s_{eo}^2 \tag{44}$$

The expression for the estimate of $\mathbf{S}_\alpha^2$ given in (43) can also be derived directly from (38) by putting $p = 1$ and substituting for s_u^2 from (42). It follows from (25) that for N and M infinitely large and $p = 1$,

$$\text{Est. } V(\bar{y}_{..}) = \frac{s_e^2}{m} \tag{45}$$

which, using (42), can also be put in the alternative form

$$\text{Est. } V(\bar{y}_{..}) = \frac{1}{h}\, s_u^2 + \frac{h-m}{m-1} \cdot \frac{1}{h}(s_u^2 - s_{eo}^2) \tag{46}$$

Equation (46) shows that s_u^2/h no longer gives an unbiased estimate of the variance of the estimated mean but that it is inflated by a component

$$\therefore \quad \frac{h-m}{m-1}\,\frac{1}{h}\,(s_u^2 - s_{eo}^2)$$

the latter vanishing when the differential biases are absent.

Example 10.1

This example is taken from the crop survey for estimation of the average yield of wheat conducted in Sind (Pakistan) in 1945–46. The design of the survey was stratified multi-stage sampling with subdivisions as the strata, a village as the first-stage unit, a field as the second-stage unit of sampling and a plot of 1/40 acre as the ultimate unit of sampling. Within each stratum the work was divided into two independent samples, one to be carried out by an official of the Department of Revenue, and the other by an official of the Department of Agriculture. Table 10.1 shows the estimates of the average yield for the two samples, together with the pooled analysis of variance of the whole sample. For one stratum, namely subdivision Kambar,

TABLE 10.1

Yield Survey on Wheat : Subdivision Kambar, Sind (Pakistan)

Estimates of Average Yield and Analysis of Variance

	Revenue	Agriculture	Combined
Mean yield in oz./plot	100·3	54·9	85·6
Number of experiments	25	12	37

Analysis of Variance (Oz./Plot)²

Source	D.F.	Mean Square
Between enumerators	1	16714·6 (= E)
Between villages within enumerators	15	7377·8 (= B)
Within villages	20	315·4 (= W)

10*a*.5 The Mean and Variance of a Stratified Sample in which Enumerators are Assigned the Units in their Respective Strata

The method of assigning enumerators considered in the previous sections, in which the units in the sample are randomly distributed among the different enumerators, is not common in practice, owing to the large travel costs it involves. The more common method is to assign neighbouring units in the sample falling in specified geographical areas, to two or more enumerators as may be needed, in the form of replicated sub-samples. As an example we may mention the assignment practice followed in crop surveys in India and described in Example 10.1. The design of the crop survey is a stratified sample with geographical divisions forming the strata and the villages selected from any stratum are randomly distributed among the requisite number of enumerators drawn locally. We shall first consider the case of a stratified sample of h units drawn from the k geographic strata of the population, with h_t units drawn from the t-th stratum, so that $\sum_{t=1}^{k} h_t = h$. We shall suppose that within any stratum an enumerator can enumerate a sample of $\bar{n}$ units during the period of the survey, so that m_t enumerators will be needed to enumerate h_t units in the t-th stratum, where $\bar{n}m_t = h_t$. We shall suppose that the h_t units in the t-th stratum are randomly distributed among the m_t enumerators and further suppose, for the sake of simplicity in the discussion, that the number of units in the population in the t-th stratum, N_t, and also the population of potential enumerators, M_t, are infinitely large. It will be seen that we have assumed that $p = 1$; consequently, we cannot separately evaluate S_x^2 and S_ϵ^2.

We shall use the same notation as in the previous sections except for the introduction of the letter t to indicate the t-th stratum. We shall denote by

x_i^t $(i = 1, 2, \ldots, h_t)$ the true value of the i-th unit in the sample of h_t units drawn from the t-th stratum,

and

y_{ij}^t the value reported on x_i^t by the j-th enumerator out of m_t in the stratum.

Clearly, the sample mean for the t-th stratum will be given by

$$\bar{y}_{..}^{t} = \frac{1}{h_t} \sum_{i}^{h_t} x_i^t + \frac{1}{m_t} \sum_{j}^{m_t} \alpha_j^t + \bar{\epsilon}_{..}^{t} \tag{47}$$

while the sample mean for the population will be given by

$$\bar{y}_{..} = \sum_{t=1}^{k} p_t \, \bar{y}_{..}^{t}$$

where

$$p_t = \frac{N_t}{N} \tag{48}$$

It follows from the results of the previous sections that

$$E\,(\bar{y}_{..}^{t} \mid h_t) = \mu_t + \bar{\alpha}_t \tag{49}$$

where μ_t = the population mean of true values in the t-th stratum, and $\bar{\alpha}_t$ = the population mean of the biases of enumerators in the t-th stratum. Also

$$\begin{aligned} E\,(\bar{y}_{..} \mid h_1, h_2, \ldots, h_k) &= \sum_{t=1}^{k} p_t \, E\,(\bar{y}_{..}^{t} \mid h_t) \\ &= \sum_{t=1}^{k} p_t \mu_t + \sum_{t=1}^{k} p_t \, \bar{\alpha}_t \\ &= \mu + \bar{\alpha} \end{aligned} \tag{50}$$

Similarly

$$V\,(\bar{y}_{..}^{t} \mid h_t) = \frac{\mathbf{S}_{tx}^2}{h_t} + \frac{\mathbf{S}_{\epsilon}^2}{h_t} + \frac{\mathbf{S}_{t\alpha}^2}{m_t} \tag{51}$$

where $\mathbf{S}_{tx}^2$ is the variance of x_i^t, $\mathbf{S}_{t\alpha}^2$ the variance of α_t and $\mathbf{S}_{\epsilon}^2$, for the sake of simplicity, is assumed to be constant from stratum to stratum; and

$$\begin{aligned} V\,(\bar{y}_{..} \mid h_1, h_2, \ldots, h_k) &= \sum_{t=1}^{k} p_t^2 \, V\,(\bar{y}_{..}^{t} \mid h_t) \\ &= \sum_{t=1}^{k} p_t^2 \left\{ \frac{\mathbf{S}_{tx}^2 + \mathbf{S}_{\epsilon}^2}{h_t} + \frac{\mathbf{S}_{t\alpha}^2}{m_t} \right\} \end{aligned} \tag{52}$$

The estimates of $S_{tx}^2 + S_\epsilon^2$ and of S_{ta}^2 are provided by the same formulæ as before. We have

$$\text{Est. } (S_{tx}^2 + S_\epsilon^2) = s^2_{t(eo)} \tag{53}$$

and

$$\text{Est. } S_{ta}^2 = s_{te}^2 - \frac{m_t}{h_t} s^2_{t(eo)} \tag{54}$$

whence

$$\text{Est. } V(\bar{y}_{..}^t \mid h_t) = \frac{s_{te}^2}{m_t} \tag{55}$$

and

$$\text{Est. } V(\bar{y}_{..} \mid h_1, h_2, \ldots, h_k) = \sum_{t=1}^{k} p_t^2 \frac{s_{te}^2}{m_t} \tag{56}$$

Example 10.2

The data for this example are derived from a pilot survey for comparing the relative efficiency of plots of different size in estimating the average yield of irrigated wheat in Moradabad District of U.P. (India) in 1944–45. The design of the survey and the method of assigning enumerators were similar to those described in Example 10.1. Thus, in each of the five subdivisions of the district, two independent samples of two villages each were selected and allocated to two enumerators designated A and B. In each village two fields were selected and in each field two plots of each size were marked. The data relating to the plot size of an equilateral triangle of side 25 links (117·9 sq.ft.) are taken here for illustration. Table 10.2 gives estimates of the average yield together with the analysis of variance for individual subdivisions, and also the pooled values for the district. Estimate the contribution to the total variation due to S_a^2.

The calculations are straightforward. Substituting from the table the values of E and B in the formula

$$\text{Est. } S_{ta}^2 = s_{te}^2 - \frac{m_t}{h_t} s^2_{t(eo)}$$

$$= \frac{(E_t - B_t)}{8}$$

TABLE 10.2

Yield Survey on Wheat in Moradabad District

Estimates of Average Yield of Wheat (Oz./Plot)

Subdivision	1		2		3		4		5		Total for the District	
	No. of Plots	Average	No. of Plots	Average	No. of Plots	Average	No. of Plots	Average	No. of Plots	Average	No. of Plots	Average
A	8	429·1	8	326·0	8	220·5	8	451·0	8	586·2	40	402·6
B	8	530·0	8	207·8	6	367·2	8	304·0	8	348·7	38	350·7
Pooled	16	479·6	16	266·9	14	283·4	16	377·5	16	467·5	78	377·3

Analysis of Variance (Oz./Plot)^2

	D.F.	Mean Square	D.F.	Mean Square	D.F.	Mean Square	D.F.	Mean Square	D.F.	Mean Square	D.F.	Mean Square
Between Subdivisions	..	..	..	..	..	..	..	..	..	..	4	154037·9
Between Enumerators	1	40703·1	1	55932·2	1	73752·3	·1	86436·0	1	225625·0	5	96489·7
Between Villages ..	2	13076·6	2	40006·2	2	30802·0	2	128254·2	2	49356·5	10	52299·1
S_a^2	..	3453·3	..	1990·7	..	5368·8	..	..	..	22033·6	..	5523·8

we obtain the values shown in the last row of the table. The average magnitude of $S_{t\alpha}^2$ over the five subdivisions works out to 6569·3.

Also

$$\text{Est.}\ (S_{tx}^2 + S_\epsilon^2) = \frac{B_t}{4}$$

whence the average value of $\hat{S}_{tx}^2 + \hat{S}_\epsilon^2$ for the district works out to 13074·8. The total variance of an observation is given by

$$\begin{aligned} \hat{S}_y^2 &= \hat{\hat{S}}_\alpha^2 + \hat{\hat{S}}_x^2 + \hat{\hat{S}}_\epsilon^2 \\ &= 19644{\cdot}1 \end{aligned}$$

Thus $\hat{\hat{S}}_\alpha^2$ accounts for about 33% of the total variation.

It should be pointed out that both in this and the previous examples the estimates of the component of variance due to differential bias are based on small numbers and, therefore, not sufficiently reliable. They are presented here only for purposes of illustration.

10*a*.6 The Mean and Variance of an Unstratified Sample in which Enumerators are Assigned Neighbouring Units

Sometimes p_t is not known and yet we wish to assign to the enumerators neighbouring units in the sample falling in specified geographical areas, in the manner indicated above, after an unstratified sample of h has been selected from the population. In this situation the number of units falling in the t-th stratum, namely h_t, is a random variable, and so is m_t, while the estimate of p_t is provided by h_t/h.

The sample mean $\bar{y}_{..}$ for this case is given by

$$\bar{y}_{..} = \frac{1}{h} \sum_{t=1}^{k} h_t \bar{y}_{..}^{t} \tag{57}$$

and its expected value and variance for a given set of $h_1, h_2, \ldots, h_k$ by

$$E(\bar{y}_{..} \mid h_1, h_2, \ldots, h_k) = \frac{1}{h} \sum_{t=1}^{k} h_t (\mu_t + \bar{a}_t) \tag{58}$$

and

$$V(\bar{y}_{..} \mid h_1, h_2, \ldots, h_k) = \sum_{t=1}^{k} \frac{h_t^2}{h^2} \left\{ \frac{\mathbf{S}_{tx}^2 + \mathbf{S}_\epsilon^2}{h_t} + \frac{\mathbf{S}_{ta}^2}{m_t} \right\}$$

$$= \frac{1}{h^2} \sum_{t=1}^{k} h_t (\mathbf{S}_{tx}^2 + \mathbf{S}_\epsilon^2) + \frac{\bar{n}}{h^2} \sum_{t=1}^{k} h_t \, \mathbf{S}_{ta}^2 \tag{59}$$

It is seen from (58) that the conditional expected value of $\bar{y}_{..}$ does not equal $\mu + \bar{a}$. To the expression (59) we must, therefore, add the square of the bias component and then take the expectation in order to obtain the variance of $\bar{y}_{..}$. We write

$$V(\bar{y}_{..}) = E\{V(\bar{y}_{..} \mid h_1, h_2, \ldots, h_k)\} + E\{E(\bar{y}_{..} \mid h_1, h_2, \ldots, h_k) - \mu - \bar{a}\}^2$$

$$= E\left\{ \frac{1}{h^2} \sum_{t=1}^{k} h_t (\mathbf{S}_{tx}^2 + \mathbf{S}_\epsilon^2) + \frac{\bar{n}}{h^2} \sum_{t=1}^{k} h_t \mathbf{S}_{ta}^2 \right\}$$

$$+ E\left\{ \sum_{t=1}^{k} \left(\frac{h_t}{h} - p_t \right) (\mu_t + \bar{a}_t) \right\}^2 \tag{60}$$

The value of the first part in (60) is clearly given by

$$\frac{1}{h} \sum_{t=1}^{k} p_t (\mathbf{S}_{tx}^2 + \mathbf{S}_\epsilon^2) + \frac{\bar{n}}{h} \sum_{t=1}^{k} p_t \mathbf{S}_{ta}^2 \tag{61}$$

while for the second, we write

$$E\left\{ \sum_{t=1}^{k} \left(\frac{h_t}{h} - p_t \right) (\mu_t + \bar{a}_t) \right\}^2$$

$$= E\left\{ \sum_{t=1}^{k} \left(\frac{h_t}{h} - p_t \right)^2 (\mu_t + \bar{a}_t)^2 \right.$$

$$\left. + \sum_{t \neq t'=1}^{k} \left(\frac{h_t}{h} - p_t \right) \left(\frac{h_{t'}}{h} - p_{t'} \right) (\mu_t + \bar{a}_t)(\mu_{t'} + \bar{a}_{t'}) \right\}$$

$$= \sum_{t=1}^{k} \frac{N-h}{N-1} \cdot \frac{p_t(1-p_t)}{h} (\mu_t + \bar{a}_t)^2$$

$$- \sum_{t \neq t'=1}^{k} \frac{N-h}{N-1} \cdot \frac{p_t p_{t'}}{h} (\mu_t + \bar{a}_t)(\mu_{t'} + \bar{a}_{t'})$$

$$= \frac{N-h}{N-1} \cdot \frac{1}{h} \left\{ \sum_{t=1}^{k} p_t (\mu_t + \bar{a}_t)^2 - \sum_{t=1}^{k} p_t^2 (\mu_t + \bar{a}_t)^2 \right.$$

$$\left. - \sum_{t \neq t'=1}^{k} p_t p_{t'} (\mu_t + \bar{a}_t)(\mu_{t'} + \bar{a}_{t'}) \right\}$$

$$= \frac{N-h}{N-1} \cdot \frac{1}{h} \left[\sum_{t=1}^{k} p_t (\mu_t + \bar{a}_t)^2 - \left\{ \sum_{t=1}^{k} p_t (\mu_t + \bar{a}_t) \right\}^2 \right]$$

$$= \frac{N-h}{N-1} \cdot \frac{1}{h} \left[\sum_{t=1}^{k} p_t (\mu_t + \bar{a}_t - \mu - \bar{a})^2 \right]$$

$$\cong \frac{1}{h} \sum_{t=1}^{k} p_t (\mu_t + \bar{a}_t - \mu - \bar{a})^2 \tag{62}$$

Substituting from (61) and (62) in (60), we get

$$V(\bar{y}_{..}) = \frac{1}{h} \sum_{t=1}^{k} p_t (S_{tx}^2 + S_\epsilon^2) + \frac{\bar{n}}{h} \sum_{t=1}^{k} p_t S_{t\alpha}^2$$

$$+ \frac{1}{h} \sum_{t=1}^{k} p_t (\mu_t + \bar{a}_t - \mu - \bar{a})^2 \tag{63}$$

The expression (63) can be simplified. Adding and subtracting from it $(1/h) \sum_{t=1}^{k} p_t S_{t\alpha}^2$, we may write

$$V(\bar{y}_{..}) = \frac{1}{h}\left\{\sum_{t=1}^{k} p_t\,(\mathbf{S}_{tx}^2 + \mathbf{S}_\epsilon^2 + \mathbf{S}_{ta}^2)\right\}$$

$$+ \frac{1}{h}\sum_{t=1}^{k} p_t\,(\mu_t + \bar{a}_t - \mu - \bar{a})^2$$

$$+ \frac{\bar{n}}{h}\sum_{t=1}^{k} p_t \mathbf{S}_{ta}^2 - \frac{1}{h}\sum_{t=1}^{k} p_t \mathbf{S}_{ta}^2 \qquad (64)$$

Further, we note that the variance of an observation can be expressed as

$$\mathbf{S}_y^2 = \sum_{t=1}^{k} p_t\,(\mathbf{S}_{tx}^2 + \mathbf{S}_\epsilon^2 + \mathbf{S}_{ta}^2) + \sum_{t=1}^{k} p_t\,(\mu_t + \bar{a}_t - \mu - \bar{a})^2 \qquad (65)$$

whence

$$V(\bar{y}_{..}) = \frac{\mathbf{S}_y^2}{h} + \left(\frac{1}{m} - \frac{1}{h}\right)\sum_{t=1}^{k} p_t \mathbf{S}_{ta}^2 \qquad (66)$$

If we define $\mathbf{S}_a^2 = \sum_{t=1}^{k} p_t \mathbf{S}_{ta}^2$, we obtain for $V(\bar{y}_{..})$ the same expression as in (26), as we would in fact expect.

10*a*.7 An Alternative Expression for the Variance of the Sample Mean in Terms of the Covariance between Responses Obtained by the Same Enumerator

Let $\mathbf{S}_{yI}$ denote the covariance between observations recorded by the same enumerator. Then, clearly

$$\mathbf{S}_{yI} = E\,E\left\{\Big(y_{ij} - E\,(y_{ij})\Big)\Big(y_{i'j} - E\,(y_{i'j})\Big) \Big|\ j\right\}$$

$$= E\,E\,\{(x_i + a_j + \epsilon_{ij} - \mu - \bar{a})\,(x_{i'} + a_j + \epsilon_{i'j} - \mu - \bar{a}) \mid j\}$$

$$= E\,E\,[(x_i - \mu)\,(x_{i'} - \mu) + (a_j - \bar{a})^2 + \epsilon_{ij}\,\epsilon_{i'j}$$

$$+ (a_j - \bar{a})\,\{(x_i - \mu) + (x_{i'} - \mu)\}$$

$$+ (a_j - \bar{a})\,(\epsilon_{ij} + \epsilon_{i'j})$$

$$+ \epsilon_{ij}\,(x_{i'} - \mu) + \epsilon_{i'j}\,(x_i - \mu)]$$

$$= -\frac{\mathbf{S}_\epsilon^2}{N} + \mathbf{S}_a^2\left(1 - \frac{1}{M}\right)$$

since all other terms are zero,

$$\cong \mathbf{S}_{\alpha}^{2} \tag{67}$$

for large N and M. Substituting in (26), we get

$$V(\bar{y}_{..}) = \frac{1}{h}\,\mathbf{S}_{y}^{2} + \mathbf{S}_{yI}\left[\frac{1}{m} - \frac{1}{h}\right] \tag{68}$$

The formula in this form was first derived by Hansen, Hurwitz, Marks and Mauldin (1951). It is instructive since it shows that if errors of observation are unrelated from unit to unit so that $\mathbf{S}_{yI} = 0$, the variance is given by the usual expression for samples from a large population. It follows also that when an enumerator is given only one unit to enumerate, in which case both $m = h$ and $\mathbf{S}_{yI} = 0$, the variance takes the minimum value, being given by the usual expression $\mathbf{S}_{y}^{2}/h$.

10*a*.8 Determination of the Optimum Number of Enumerators

We have seen that the variance of the sample mean is decreased as the number of enumerators participating in the survey is increased. Practical considerations, however, place a limit on the number to which enumerators can be increased. Obviously, the principle to be followed in determining this number is the principle of minimizing the variance of the sample mean for a given cost allotted to the survey.

When a sample of h units is randomly and equally distributed among m enumerators, then it is reasonable to suppose that the cost of the survey will be represented by

$$C = c_1 h + c_2 m + c_3 \sqrt{hm} \tag{69}$$

where

c_1 is the cost of collecting information per unit,

c_2 the cost of engaging an enumerator,

and

c_3 proportional to the cost of travel on unit distance.

We shall further suppose that the money allotted to the survey is fixed at C_0. To calculate the optimum values of h and m, we use the method of Lagrangian multipliers and form a function ϕ given by

$$\phi = \frac{S_y^2}{h} + S_\alpha^2\left[\frac{1}{m} - \frac{1}{h}\right] + \mu\,(c_1h + c_2m + c_3\sqrt{hm} - C_0) \qquad (70)$$

where μ is a Lagrangian constant.

Differentiating ϕ with respect to h, m and μ and equating to zero, we obtain

$$\frac{\partial\phi}{\partial h} = -\frac{S_y^2}{h^2} + \frac{S_\alpha^2}{h^2} + \mu\left[c_1 + \frac{c_3}{2}\sqrt{\frac{m}{h}}\right] = 0 \qquad (71)$$

$$\frac{\partial\phi}{\partial m} = -\frac{S_\alpha^2}{m^2} + \mu\left[c_2 + \frac{c_3}{2}\sqrt{\frac{h}{m}}\right] = 0 \qquad (72)$$

and

$$\frac{\partial\phi}{\partial\mu} = c_1h + c_2m + c_3\sqrt{hm} - C_0 = 0 \qquad (73)$$

whence, multiplying both sides of (71) by h and of (72) by m and eliminating μ, we obtain

$$x^4 + \frac{c_3}{2c_2}x^3 - \frac{c_3}{2c_1}\beta^2x - \beta^2 = 0 \qquad (74)$$

where

$$x^2 = \frac{m}{h} \quad \text{and} \quad \beta^2 = \frac{c_1}{c_2}\cdot\frac{S_\alpha^2}{S_y^2 - S_\alpha^2} \qquad (75)$$

An inspection of (74) shows that the equation has two real roots, one positive and one negative. An explicit expression for the roots is, however, difficult to obtain and the solution has therefore to be reached by the trial and error method.

Two special cases of the cost function are of interest. When c_1 is zero, the equation (74) reduces to a cubic

$$x^3 + \frac{c_3}{2c_2}x^2 - \frac{c_3}{2c_2}\cdot\frac{S_\alpha^2}{S_y^2 - S_\alpha^2} = 0 \qquad (76)$$

but here again an explicit expression for x is difficult to obtain.

The other case of interest is when $c_3 = 0$. We get

$$\frac{m^2}{h^2} = \beta^2 \tag{77}$$

Substituting for m from (77) in (73) after putting $c_3 = 0$, we get

$$c_1 h + c_2 \beta h = C_0 \tag{78}$$

or

$$h = \frac{C_0}{c_1 + c_2\beta} \tag{79}$$

and

$$m = \frac{\beta C_0}{c_1 + c_2\beta} \tag{80}$$

Equation (80) shows, what is of course obvious, that the larger the contribution of S_a^2 relative to S_y^2, the larger should be the number of enumerators participating in the survey. On the other hand, it is likely that as the number of enumerators is increased, the difficulties of controlling the survey also increase, possibly resulting in a larger value of $\bar{a}$.

We shall not illustrate here the application of this result but shall refer to two more examples to gain some idea of the contribution of S_a^2 relative to the total variation and the importance of controlling it in surveys.

Example 10.3

This relates to a socio-economic survey conducted by students of the International Training Centre on Censuses and Statistics for South-East Asia during December 1949. The survey was carried out in three villages: Badli, Shamapur and Auchandi, situated near Delhi (India). The houses in each village were serially numbered and grouped into blocks of three. A certain number of these blocks was selected at random and within each block alternate households were enumerated. The sample for each village was divided into independent samples, one each to be enumerated by a different party of students. Thus, the work in the village Badli was divided among six parties of enumerators, that in Shamapur and Auchandi among four and two parties respectively. The questionnaire used for the survey was prepared by the students themselves and included a large number of items. The results given here relate only to two characters, *viz.*, the proportion of illiterates and the proportion of persons economically independent in a household. Table 10.3 gives the estimated values for each of the two characters for one village, Badli. The table shows that there is more variability in the estimates given by different parties in the character "economic independence" than in the other character. The estimated values of S_α^2 and of the total variation, *i.e.*, $S_x^2 + S_\alpha^2 + S_\epsilon^2$ are also given in Table 10.3. The relative magnitude of S_α^2 as compared to the total variation is seen to be larger in the case of "economic independence" than in the case of the other character as expected, but not significantly so, being based on a small number of cases.

TABLE 10.3

Socio-Economic Survey in Badli

Party	A	B	C	D	E	F
% Illiterate	88	75	90	87	62	95
% Economically independent	67	46	77	50	46	31

	Values of $\hat{S}_\alpha^2$ *and* $\hat{S}_x^2 + \hat{S}_\alpha^2 + \hat{S}_\epsilon^2$	
% Illiterate	0	574
% Economically independent	104	722

Example 10.4

This relates to the data on acreage collected in the course of a surprise check to which a reference has already been made in Section 1.8. Acreage under crops in India is compiled by the village accountants by noting the names of the crops field by field in the course of their administrative duties. As all the fields are surveyed and mapped and the area of each field (survey number) is, therefore, accurately known, the total area under any crop is obtained by simply adding the area of the fields growing that crop.

This method, though sound in principle, is not always free from errors in practice. In fact, there is a criticism that the village accountants do not exercise sufficient care in ascertaining the names of crops grown in the fields in their respective villages. A surprise check, was, therefore, organized in randomly selected villages of Lucknow District in Uttar Pradesh (India) to examine the extent of the inaccuracy of the records maintained by the village accountants. The check was carried out by the Statistical Staff of the Department of Agriculture, Uttar Pradesh, and of the Indian Council of Agricultural Research, New Delhi.

Altogether 61 villages were selected for the purpose of this check. In each village 8 fields were selected at random. In each field the statistical investigator was asked to record the name of the crop. In case more than one crop was grown, he was asked to give the names of the crops together with the proportion of the area under each. The village accountants' records for the same fields were taken from the register maintained by them. The check was carried out at harvest time after the village accountants had completed their inspection and made entries in the register. Table 10.4 summarises the results.

TABLE 10.4

Comparison of Crop-Acreages

Name of Crop	Gross Area (Sq.ft.) as recorded by		Difference (2) – (3)	Percentage over (3)
	Village Accountants	Statistical Staff		
(1)	(2)	(3)	(4)	(5)
Wheat ..	3313224	3227149	86075	+ 2·7
Gram ..	2391840	2415085	– 23245	– 1·0
Barley ..	1873291	2029875	–156584	– 7·7
Arhar ..	1888117	2103171	–215054	–10·2

It is seen that the discrepancy varies from + 2·7% in wheat to – 10·2% in *arhar*. Although the discrepancies do not appear to be serious or in the same direction, they point to the need for strengthening the supervision over the work of the field staff and the conduct of similar checks in other parts of the country.

This conclusion is confirmed by treating the data by the methods developed in this chapter. We have

$$k = 61, \; n_{ij}^{t} = 0 \text{ or } 1, \; n_{i.}^{t} = p_{.} = 2, \; h_t = 8, \; m = 2$$

On analysing the data, it was found that S_α^2 did not exceed 5% of the total variation in the case of any crop.

10*a*.9 Limitations of the Method of Replicated Samples in Surveys

We have seen that when a survey is arranged in the form of replicated samples, each one to be enumerated by a different enumerator, we can estimate S_α^2 though we cannot separate μ from $\bar{\alpha}$. We also saw that S_α^2 may be considerable, depending upon the character observed and the conditions of the survey. In general, it may be said that for characters whose value is influenced by the judgment of the enumerator, like an eye-estimate of the yield of a crop, or for those for which it is difficult to obtain a precise answer, like farm expenditure, S_α^2 may be considerable. S_α^2 measures, as it were, the influence of enumerators on response and is likely to be particularly large in surveys in underdeveloped countries.

A question naturally arises whether a sample should be randomly divided among enumerators participating in the survey as a matter of course in order to make possible the measurement of S_α^2 and test its significance relative to the total variation.

Obviously, the answer must depend upon the considerations of cost and of the precision with which it is possible to test the significance of S_α^2. For any single stratum the test of significance of S_α^2 is provided by s_e^2/s_{eo}^2, but it does not have a good discriminating power as the size of the sample allotted to any single stratum is normally small. In consequence, the test would most of the time fail to reveal the significant existence of S_α^2 even when the differences between the enumerators' estimates are large.

One method of improving the discriminating power of the test is to use what are termed *linked samples*. In this arrangement, linked pairs (or groups) of sampling units are selected and one sampling unit of each pair is allotted to one enumerator and the other to the second enumerator (Mahalanobis, 1944). This helps to reduce the variance of the difference between the estimates. On the other hand, it is easy to see from Section 7.6 that the variance of the pooled estimate is increased by $S_y^2\,(m-1)\,\rho/n$, where ρ is the intra-class correlation between the m members of linked samples, m is the number of enumerators and n is the total number of units in the sample. There is thus a loss in efficiency of the combined sample estimate in relation to the procedure where linked samples are not used. Any decision to use linked samples has therefore to be made after ascertaining the increase in expenditure on the survey for attaining the desired precision.

Another method of overcoming the difficulty is to pool the results over all the strata and use s_{te}^2/s_{eo}^2 to test for differential bias among enumerators. This test is more sensitive than the one for a single stratum but is of limited use as it does not help to locate the disagreement among enumerators, since for the location of disagreement the results of individual strata must be examined, but as stated in the previous paragraph the test of significance for individual strata will not reveal the differential bias most of the time.

Apart from the limitation of the size of the samples within a stratum which renders replicated samples an ineffective tool for detecting discrepancies in fieldwork, it should be mentioned that the use of the replicated samples requires that each enumerator cover the entire stratum thereby adding to the travel component in the cost of the survey. Assuming that m' enumerators, as against m in replicated samples, are required in case the sample is divided into m' compact groups, one such group to be enumerated by one enumerator, the cost of the survey will be increased roughly in proportion to $(\sqrt{m}-1)$. This is on the assumption that the travelling cost per enumerator within an area A is proportional to $\sqrt{An'}$, where n' is the number of units to be enumerated by each enumerator distributed randomly over the area. If travel cost forms a considerable part of the cost of a survey, the loss in efficiency resulting from replication may be appreciably large.

These are the limitations one should bear in mind in using the method of replicated samples in surveys. Further, in interpreting the results of replicated samples, there is a danger of guiding oneself into the belief that the fieldwork is under control, when in actual fact it may well be otherwise. An example will help to illustrate the point.

This example is taken from a jute survey in Bengal (Mahalanobis, 1944). The object of the survey was to estimate the area under jute. The survey was organized in the form of two samples *A* and *B*. In each of the 379 strata into which Bengal was divided for purposes of the survey, the difference between the estimates *A* and *B* was tabulated and tested for significance. Table 10.5 reproduced below gives the distribution of the values of *t* and shows that in 109 out of 379 cases, the value of *t* was significant, although the expected number of significant cases at this level is only 19.

TABLE 10.5

Comparison of Samples A and B; Student's t for Strata

Range of Probability of *t*-Values	Number of Cases		Difference	χ^2
	Observed	Expected		
Less than 0·05	109	18·95	+90·05	427·92
0·05–0·10	20	18·95	+ 1·05	0·06
0·10–0·90	235	303·20	−68·20	15·34
0·90–0·95	12	18·95	− 6·95	2·55
0·95–1·00	3	18·95	−15·95	13·42
Total ..	379	379·00		459·29

In order to find an explanation for this large difference a scrutiny of the field records was made by the author of the survey and showed that in 84 out of the 109 strata, the discrepancy could be ascribed to the influence of real physical differences, such as weather conditions during the periods in which the work of the *A* and *B* samples was carried out. Omitting these 84 cases, there are 295 left. The distribution of the remaining 295 values of *t* was found to be in satisfactory agreement with the expected distribution of t. It was concluded that "the object of using the replicated sampling method was entirely successful".

This interpretation, however, raises a logical difficulty. Once discrepant work is suspected, it would appear only proper to scrutinize the work in all the strata and not confine the scrutiny to only those having significant values of t. For, in a stratum where t is non-significant, one can also expect discrepant work since the non-significance can be due to the opposite effects of the discrepancy in work and the real physical differences between the two samples A and B. When the sample size is small, as it will usually be in each stratum, this method may lead one to looking for trouble where it does not exist and *vice versa*, as it is likely that real large differences may be declared non-significant and *vice versa*. There also arise practical difficulties in going back to the sampling units for the scrutiny required after the survey is over. The whole procedure of accepting the verdict of agreement where t is non-significant and explaining the difference in terms of physical differences where t is significant, is logically untenable.

Even apart from considerations of cost and interpretation, a random allotment of the sampling units among enumerators cannot by itself be an effective tool for the control of fieldwork, and the need of controlling it in other ways is obvious. This need it would appear is best met by providing adequate and effective supervision over fieldwork. We shall conclude this discussion by examining the roles of supervision and replication at the primary level. The two differ in several respects:

(i) Supervision is carried out by the superior staff, better paid, qualified and experienced as compared to the enumerators at the primary level employed in replicated samples.

(ii) It is carried out on a part of the work performed at the primary level whereas replicated samples require at least two independent samples.

(iii) Supervision is not confined only to enumerating for the second time units once observed at the primary level. It has a wider objective in view, namely that of correcting and improving the fieldwork on the spot, whereas replicated samples will usually suggest the need for improvement when the survey is over.

(iv) A supervisor need not be present throughout the operations connected with the enumeration of a selected unit, whereas an enumerator under sample survey must enumerate completely every unit assigned to him.

(v) Units selected for supervision may or may not be selected by the principle of random sampling, whereas in replicated samples they will necessarily be so selected. When it is possible to arrange supervision on a probability basis and the work done by the supervisors is considered a sub-sample of the work done at the primary level, supervision may be considered a very special form of replicated samples subject to the differences mentioned above. This way supervision can be utilised to improve the estimates obtained from the work done at the primary level.

(vi) Replicated samples will not reveal minor defects in an investigator and will certainly not reveal faults which are common to all the investigators, whereas this is possible with supervisory checks.

(vii) Replicated samples alone can estimate observational errors whereas supervision will not, unless conducted as visualised in (v).

It would be seen that supervision can provide a better control over fieldwork in a variety of ways which is not possible in the case of replicated samples. Replicated samples are no alternative to supervisory check, though the latter can be. Replicated samples have a place either when the object of the survey is to compare different methods or different classes of investigators, or at the pilot stage of a large-scale survey for testing questionnaires and procedures, but would hardly appear worth while for adoption as a regular feature of surveys.

B. INCOMPLETE SAMPLES

10*b*.1 The Problem

It is common experience that some of the units selected in the sample do not respond, at least at the first attempt, and indeed may not respond even after repeated attempts. Thus the selected farmers or families may not be found at home at the first attempt and some may refuse to co-operate with the interviewer even if contacted at the second attempt. Persuasion and further attempts are therefore invariably required for achieving completeness. This, however, increases the cost of the survey. On the other hand, estimates based on the incomplete sample may be biased. This extent of incompleteness, called *non-response*, is sometimes so large as to completely vitiate the estimate. The problem is particularly important in interview surveys. In the following section, we shall give the solution of the problem as first put forward by Hansen and Hurwitz (1946), which consists in drawing a sub-sample of non-respondents and enumerating it completely through later attempts, the size of the total sample and that of the sub-sample in the non-response group being so determined as to give an unbiased estimate of the population value with the desired precision at minimum cost.

10*b*.2 The Solution of the Problem of Incomplete Samples

We shall suppose that the population can be divided into two classes, those who will respond at the first attempt and those who will not. For convenience we shall call the two classes as the *response* and *non-response classes*. If n_1 units in the sample respond and n_2 do not, then we may regard n_1 a random sample of the response class and n_2 a random sample of the non-response class. Let h_2 denote the size of the sub-sample from n_2 to be enumerated at the second attempt, such that

$$n_2 = fh_2 \qquad (81)$$

Further, let N_1 and N_2 denote the sizes of the response and the non-response classes in the population. Clearly, N_1 and N_2 cannot be known and can only be estimated from the sample. We have

$$\text{Est. } N_1 = \frac{n_1 N}{n}$$

and

$$\text{Est. } N_2 = \frac{n_2 N}{n}$$

The cost of the survey will be made up of three parts as follows:

$$C = c_0 n + c_1 n_1 + c_2 h_2 \tag{82}$$

where c_0 represents the cost of locating a sample unit at the first attempt, c_1 the cost of enumerating and processing information per unit in the response class, and c_2 the cost of enumerating and processing information per unit in the non-response class. This cost will obviously vary from sample to sample. We shall therefore consider the average cost of the survey. Substituting for n_1 and n_2 their expected values, we get

$$C = \frac{n}{N}\left\{N c_0 + N_1 c_1 + \frac{N_2}{f}\, c_2\right\} \tag{83}$$

Clearly, the estimate of the population mean is given by

$$\bar{y}_w = \frac{\hat{N}_1}{N}\,\bar{y}_{n_1} + \frac{\hat{N}_2}{N}\,\bar{y}_{h_2} = \frac{\{n_1\bar{y}_{n_1} + n_2\bar{y}_{h_2}\}}{n} \tag{84}$$

It is easily shown that this gives an unbiased estimate of the population mean, for,

$$\begin{aligned} E\,(n_1\bar{y}_{n_1} \mid n) &= E\,\{E\,(n_1\bar{y}_{n_1} \mid n_1, n)\} \\ &= E\,\{n_1\bar{y}_{N_1} \mid n\} \\ &= \frac{nN_1\bar{y}_{N_1}}{N} \end{aligned}$$

and

$$\begin{aligned} E\,(n_2\bar{y}_{h_2} \mid n) &= E\,E\,E\,(n_2\bar{y}_{h_2} \mid h_2, y_1, \ldots, y_{n_2}, n) \\ &= E\,E\,(n_2\bar{y}_{n_2} \mid n_2, n) \\ &= E\,(n_2\bar{y}_{N_2} \mid n) \\ &= \frac{nN_2\bar{y}_{N_2}}{N} \end{aligned}$$

whence

$$E(\bar{y}_w) \doteq \frac{\{N_1\bar{y}_{N_1} + N_2\bar{y}_{N_2}\}}{N}$$

$$= \bar{y}_N \tag{85}$$

To obtain the variance of $\bar{y}_w$, we have

$$V(\bar{y}_w) = E\left\{\frac{n_1\bar{y}_{n_1} + n_2\bar{y}_{h_2}}{n} - \bar{y}_N\right\}^2$$

The right-hand side may be written as

$$= E\left\{\frac{n_1\bar{y}_{n_1} + n_2\bar{y}_{n_2} - n_2\bar{y}_{n_2} + n_2\bar{y}_{h_2}}{n} - \bar{y}_N\right\}^2$$

$$= E\left\{\frac{n_1\bar{y}_{n_1} + n_2\bar{y}_{n_2}}{n} - \bar{y}_N + \frac{n_2\bar{y}_{h_2} - n_2\bar{y}_{n_2}}{n}\right\}^2$$

$$= E\left\{(\bar{y}_n - \bar{y}_N) + \frac{n_2}{n}(\bar{y}_{h_2} - \bar{y}_{n_2})\right\}^2$$

$$= E\left\{(\bar{y}_n - \bar{y}_N)^2 + \frac{n_2^{\,2}}{n^2}(\bar{y}_{h_2} - \bar{y}_{n_2})^2 + \frac{2n_2}{n}(\bar{y}_n - \bar{y}_N)(\bar{y}_{h_2} - \bar{y}_{n_2})\right\} \tag{86}$$

We know already that

$$E(\bar{y}_n - \bar{y}_N)^2 = \left(\frac{1}{n} - \frac{1}{N}\right) \mathrm{S}^2 \tag{87}$$

where S^2 is the mean square for the whole population.

To evaluate the second term in (86), we first have

$$E(\bar{y}_{h_2} - \bar{y}_{n_2})^2 = E\{E(\bar{y}_{h_2} - \bar{y}_{n_2})^2 \mid h_2, y_1, \ldots, y_{n_2}\}$$

$$= E\left\{\left(\frac{1}{h_2} - \frac{1}{n_2}\right) s_2^{\,2} \mid h_2, n_2\right\}$$

where

$$s_2^{\,2} = \frac{\sum^{n_2} (y_i - \bar{y}_{n_2})^2}{n_2 - 1}$$

so that

$$E(\bar{y}_{h_2} - \bar{y}_{n_2})^2 = \left(\frac{1}{h_2} - \frac{1}{n_2}\right) \mathbf{S}_2^2 \tag{88}$$

where

$$\mathbf{S}_2^2 = \frac{1}{N_2 - 1} \sum_{i=1}^{N_2} (y_i - \bar{y}_{N_2})^2$$

Hence

$$E\left\{\frac{n_2^2}{n^2}(\bar{y}_{h_2} - \bar{y}_{n_2})^2\right\} = E\left\{\frac{n_2^2}{n^2}\left(\frac{1}{h_2} - \frac{1}{n_2}\right)\mathbf{S}_2^2\right\}$$

$$= \frac{f-1}{n^2} E\{n_2 \mathbf{S}_2^2\}$$

$$= \frac{f-1}{n} \cdot \frac{N_2}{N} \mathbf{S}_2^2 \tag{89}$$

The value of the third term in (86) is clearly zero. Hence, from (86), (87) and (89), we get

$$V(\bar{y}_w) = \left(\frac{1}{n} - \frac{1}{N}\right) \mathbf{S}^2 + \frac{f-1}{n} \cdot \frac{N_2}{N} \mathbf{S}_2^2 \tag{90}$$

If $f = 1$, the second term of (90) will vanish and we shall be left with the variance of the mean of a simple random sample of n as we would expect. The second term represents the increase in the variance arising from sub-sampling h_2 out of n_2 units.

We shall now proceed to determine the optimum values of n and f. Let

$$\phi = C + \mu(V - V_0)$$

$$= \frac{n}{N}\left\{Nc_0 + N_1c_1 + \frac{N_2}{f}c_2\right\} + \mu\left\{\frac{N-n}{Nn}\mathbf{S}^2\right.$$

$$\left. + \frac{N_2}{Nn}(f-1)\mathbf{S}_2^2 - V_0\right\} \tag{91}$$

where V_0 is the value of the variance with which it is desired to estimate the mean. Differentiating with respect to n and f, and equating to zero, we obtain

$$\frac{1}{N}\left\{Nc_0 + N_1c_1 + \frac{N_2}{f}c_2\right\} = \frac{\mu}{n^2}\left\{S^2 + \frac{N_2}{N}(f-1)S_2^2\right\} \tag{92}$$

and

$$\frac{nc_2}{f^2} = \frac{\mu S_2^2}{n} \tag{93}$$

On substituting for $1/n^2$ from (93) in (92), we obtain

$$\frac{1}{N}\left\{Nc_0 + N_1c_1 + \frac{N_2}{f}c_2\right\} = \frac{c_2}{f^2S_2^2}\left\{S^2 + \frac{N_2}{N}(f-1)S_2^2\right\}$$

which reduces to

$$\frac{1}{N}\{Nc_0 + N_1c_1\} = \frac{c_2}{f^2S_2^2}\left\{S^2 - \frac{N_2}{N}S_2^2\right\}$$

Hence

$$f = \sqrt{\frac{c_2\left(S^2 - \frac{N_2}{N}S_2^2\right)}{S_2^2\left(c_0 + \frac{N_1}{N}c_1\right)}} \tag{94}$$

From (92), we obtain

$$n^2 = \frac{\mu N\left\{S^2 + \frac{N_2}{N}(f-1)S_2^2\right\}}{\left(Nc_0 + N_1c_1 + \frac{N_2}{f}c_2\right)} \tag{95}$$

Now to find μ we note that

$$V_0 = \frac{N-n}{Nn}S^2 + \frac{N_2}{N}\cdot\frac{f-1}{n}S_2^2$$

or

$$\left(V_0 + \frac{S^2}{N}\right)^2 = \frac{1}{n^2}\left\{S^2 + \frac{N_2}{N}(f-1)S_2^2\right\}^2 \tag{96}$$

On substituting for n^2 from (92), we therefore have

$$\mu = \frac{\left(Nc_0 + N_1c_1 + \frac{N_2}{f}\, c_2\right)}{N\left(V_0 + \frac{S^2}{N}\right)^2} \cdot \left\{S^2 + \frac{N_2}{N}(f-1)\, S_2^2\right\}$$

whence, on substituting the result in (95), we finally have

$$n = \frac{S^2 + \frac{N_2}{N}\,(f-1)\,S_2^2}{V_0 + \frac{S^2}{N}} \tag{97}$$

Equations (94) and (97) thus provide the values of n and f required to estimate the population mean with the desired standard error at the minimum cost.

An example will serve to illustrate the method. Suppose the response rate is 50% and S_2^2 for the non-response group is 4/5 of that in the whole population. In other words,

$$\frac{N_1}{N} = \frac{N_2}{N} = 0{\cdot}5$$

and

$$S_2^2 = \frac{4S^2}{5}$$

Ignoring the finite multiplier, we then obtain from equation (90), the variance of the estimated mean as

$$V(\bar{y}_w) = \frac{S^2(3 + 2f)}{5n} \tag{98}$$

To work out the cost of the survey let us assume that it costs one rupee to contact a unit, four rupees to enumerate and process information on that unit and eight rupees to enumerate and process information on the unit in the non-response group. The total cost of the survey is, therefore, given by

$$C = n\left(3 + \frac{4}{f}\right) \tag{99}$$

Now let us suppose that we wish to estimate the mean of the population with a desired variance V_0 equal to, say, $S^2/100$. Substituting in (98), we have

$$3 + 2f = \frac{n}{20} \tag{100}$$

Table 10.6 sets out for different values of f, the values of n obtained from the above equation and those of the cost of the survey obtained on substituting the values of n and f in (99). The expected value of h_2 is also given in the table. It will be seen from the table that the cost is the same for (i) $f = 1$, $n = 100$, and (ii) $f = 2$ and $n = 140$. For values larger than $f = 2$ and $n = 140$, the cost is higher. The most economical sample would therefore seem to lie between 100 and 140 with f between 1 and 2.

TABLE 10.6

Values of n and f which Provide Estimates of the Mean of the Same Precision

f	n	C	$E(h_2)$
1	100	700	50
2	140	700	35
3	180	780	30
4	220	880	27

The optimum values of n and f can also be directly obtained from equations (94) and (97). Thus f is found to be 1·4 and $n = 116$.

REFERENCES

1. Hansen, M. H., Hurwitz, W. N., Marks, E. S. and Mauldin, W. P. (1951) "Response Errors in Surveys," *Jour. Amer. Statist. Assoc.*, 46, 147–90.
2. Sukhatme, P. V. and Seth, G. R. (1952) . . "Non-Sampling Errors in Surveys," *Jour. Ind. Soc. Agr. Statist.*, 4, 5–41.
3. I.C.A.R., New Delhi (1947) *Report on the Crop-Cutting Survey for Estimating the Outturn of Wheat in Sind*, 1945–46.
4. ——— (1947) . . *Report on the Sampling Survey for Estimating the Outturn of Wheat in the United Provinces*, 1944–45.

5. International Training Centre on Censuses and Statistics for S.E. Asia (1950) "Report on Socio-Economic Surveys in Delhi Villages" (*Unpublished*).

6. Sukhatme, P. V. and Kishen, K. (1951) "Assessment of the Accuracy of Patwaris' Area Records," *Agriculture and Animal Husbandry*, **1**, No. 9, 36–47.

7. Mahalanobis, P. C. (1944) "On Large-Scale Sample Surveys," *Phil. Trans. Roy. Soc., London,* Series **B, 231**, 329–451.

8. Hansen, M. H. and Hurwitz, W. N. (1946) "The Problem of Non-Response in Sample Surveys," *Jour. Amer. Statist. Assoc.*, **41**, 517–29.

9. Sukhatme, P. V. (1947) .. "The Problem of Plot Size in Large-Scale Yield Surveys," *ibid.*, **42**, 297–310.

10. ———— (1953) .. "Measurement of Observational Errors in Surveys," *Revue de l' Institute International du Statistique*, **20**, No. 2.

INDEX

1103-52 Printed at The Bangalore Press, Bangalore City
by C. Vasudeva Rao, Superintendent.